High-Throughput Mapping of Diverse Functional Genomic Elements Governing BRCA2 Expression

High-throughput Mapping of Diverse Functional Genomic Elements Governing *BRCA2* Expression

Sharanya Srinivasan

A Thesis in the Field of Bioengineering and Nanotechnology

for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

May 2017

Abstract

The goal of this work is to develop a high-throughput approach to quantify the functional impact of the regulatory genome on target gene expression, and apply this system to catalogue functional *cis*-regulatory elements (CREs) that drive expression of *BRCA2*, a mutational driver of breast and ovarian cancer progression. The interactions between transcription factors and regulatory DNA modules underlie transcriptional outputs, but current techniques of *cis*-regulatory characterization utilize correlative features of enhancers, such as chromatin state, to assume CRE activity and do not measure the contributory effects of each CRE to a given target gene. Here we develop a CRISPR/Cas-based high-throughput screen to comprehensively and directly identify *cis*-regulatory sites that are necessary for *BRCA2* expression by intercalating mutations across 185 kilobases of genomic space and using a fluorescence reporter to obtain measurements of diminished *BRCA2* expression levels. We spatially map the distribution of required *cis*-regulatory sequences, and find evidence that proximal and distal elements exert controlling influences on *BRCA2* expression. Multiple statistical evaluations of individual site and regional significance enable clarification of a diversified and spatially dispersed functional regulatory architecture governing *BRCA2* transcription.

Acknowledgements

First and foremost, I would like to thank Richard Sherwood for being a wonderful boss, scientist and mentor, and for providing his guidance throughout the course of my thesis as Thesis Director of the project. Dr. Sherwood has motivated me to think critically and creatively in pursuing scientific questions, and has played an invaluable role in my academic and career development. I would like to thank fellow colleagues in the Sherwood laboratory at Harvard Medical School/Brigham and Women's Hospital, collaborators in David Gifford's laboratory at MIT, and co-authors on the *Nature Biotechnology* paper that pioneered a genetic screening application of the CRISPR/Cas platform. Lastly, I would like to thank my family for their constant support and encouragement as I pursued my Master's degree.

Table of Contents

List of Tables

List of Figures

Chapter I

Introduction

Gene expression is dynamically coordinated by a vast regulatory interactome, equipped by the combinatorial diversity of regulatory inputs to encode hierarchical properties of gene control and effectuate intricate regimes of cellular behavior (Spitz & Furlong, 2012). This regulatory connectivity is defined by the convergence of transcription factors (TFs) to dispersed binding-enriched genomic modules called *cis*-regulatory elements (CREs). Enhancers, a class of CREs, harbor clusters or small ensembles of short TF recognition sites for TF binding, serving as genomic platforms for conditional TF synergism and context-specific partnerships between cohorts of recruited TFs and co-activator proteins (Junion et al., 2012). The ability of an enhancer to regulate transcriptional output is governed by the element's structural and epigenetic organization, coupled with its mechanistic interactions with transcriptional proteins and genomic loci (Spitz & Furlong, 2012). Specifically, an enhancer's functionality in driving target gene expression is a complex superposition of enhancer features (DNA sequence, binding site syntax, and chromatin state) and interaction parameters (multiplicity of TF occupancy, direct TF-TF interaction contributions, strength of TF synergism with the basal transcriptional machinery), making it hard to delineate functional elements from the constellations of non-contributive sites. By controlling gene-specific agendas of expression, a subset of these functional *cis*-elements face an additional "burden": their

perturbation can cause gene dysregulation, destabilizing cellular processes of DNA repair and damage control to predispose a cancerous state. There is not only a need to focalize the abundance of TF-DNA interactions to the functionally relevant participants, but also a clinical imperative to understand how hereditary and acquired variation in functional genomic sites can contribute to the development of cancers, including breast and ovarian cancers.

Current high-throughput experimental methodologies of enhancer identification include genome-wide profiling of TF occupancy patterns, genomic surveillance of enhancer-associated histone modifications and molecular hallmarks, and DNaseI footprinting of bound DNA elements within gene-proximal regulatory regions (Neph et al., 2012; Sung, Guertin, Baek, & Hager, 2014). However, these large-scale approaches do not enable quantification of the functional significance of putative enhancers, as they utilize TF binding, histone marks, and chromatin accessibility as proxy, indirect measurements of enhancer activity (Sanjana et al., 2016). Due to experimental limitations, a large majority of these candidate enhancers have not been matched to a beneficiary target gene, and are therefore annotated as "functionally unaffiliated". On a clinical level, there is minimal understanding of the phenotypic relevance of functional regulatory nodes, the catalytic influence that non-coding sequences exert on disease progression, and the penetrative effect of variation within non-coding functional sites that regulate cancer-associated genes (Ward & Kellis, 2012).

The goal of the thesis is to elucidate the functional regulatory architecture that specifies expression of *BRCA2 (Breast Cancer 2)*, a DNA repair gene whose disruption is associated with breast and ovarian cancer progression (Welcsh & King, 2001). Here we

develop a high-throughput CRISPR/Cas mutagenesis approach to systematically and directly identify CREs that functionally contribute to *BRCA2* expression, probing the variegation of attributes of functional enhancer elements that govern *BRCA2* expression (Wang, Wei, Sabatini, & Lander, 2014). We hypothesize that *BRCA2* expression is regulated by a diverse set of enhancers that span a broad spatial distribution to facilitate proximal and long-range transcriptional activation, contain both predictive molecular hallmarks of regulatory function and non-canonical enhancer features, and exert differential contributions to *BRCA2* expression representative of the "activity strength" of the element. The high-throughput, exhaustive nature of the developed genomic screen enables unbiased interrogation of distal regulatory DNA regions, as well as genomic stretches that lack the stereotypical epigenetic and chromatin markers that associate with enhancer domains.

In the future, experimental results can be intersected with genome-wide association studies for hereditary breast and ovarian cancer patients to map significant non-coding variants and polymorphisms that occur in functionally relevant *BRCA2* regulatory sites. Importantly, identification of phenotypically relevant *cis*-regulatory elements enables targeted CRE genetic screening for individuals with a family history of breast and ovarian cancer, expanding the "search space" of cancer-associated genetic variants and strengthening the accuracy of hereditary cancer diagnostics (Walsh, 2015). Additionally, as cancer genome sequencing becomes more routine, acquired mutations that functionally impact BRCA2 can be pinpointed, improving phenotypic characterization of cancers.

Underlying Mechanisms of Enhancer Function

An enhancer's function in regulating transcriptional output is reliant on the enhancer's ability to act as a "regulatory junction" for an ordered, productive reaction with specific transcriptional proteins (Todeschini, Georges, & Veitia, 2014). Specifically, a regulatory element's contributive effect on target gene expression is dependent on its direct additive recruitment of specific TF combinations, as well as a "cascade interaction effect" of bound, cohabitated factors to indirectly remodel an environment for assistive TF loading or cooperatively recruit additional co-activator proteins and the basal transcriptional machinery (BTM) by direct protein-to-protein binding (He, Samee, Blatti, & Sinha, 2010). Recently, reporter assays using synthetically constructed CREs have demonstrated that certain subsets of enhancers depend on precise orientation and patterning of TF binding sites to promote enhancer activity, while other types of enhancers are structurally flexible and require little or no motif grammar to generate gene expression profiles (Erceg et al., 2014). Other studies have surveyed specific enhancers to elucidate the nature of TF co-occupancy and the mechanisms of TF cooperativity, resulting in the delineation of various modes of combinatorial regulation that include assistive alteration of the enhancer's environment to promote additional TF binding events and transcriptional synergy of bound proteins to concertedly recruit the BTM (Spitz & Furlong, 2012; Liu et al., 2014) (Figure 1).

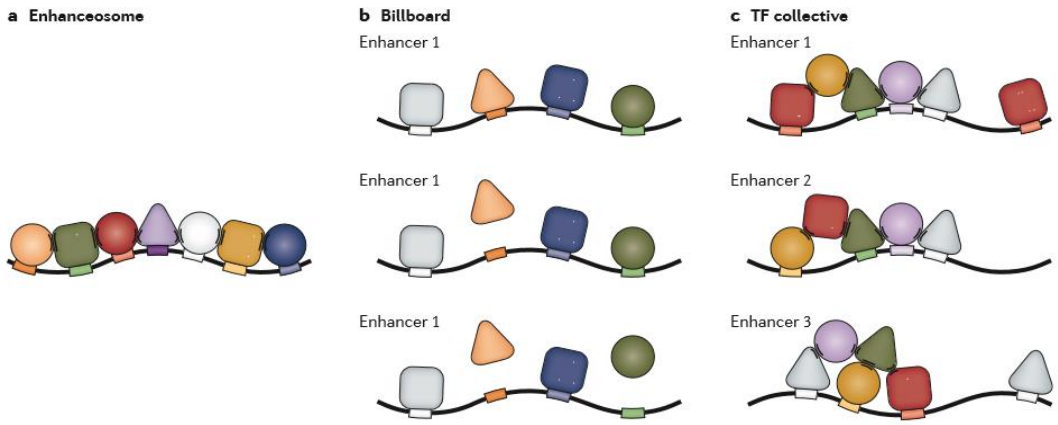

Figure 1. Three primary models of enhancer activity. Image (a) represents the enhanceosome model, which suggests that recruited TFs form an ordered configuration, and fixed motif composition and grammar is necessary for enhancer activation. Image (b) describes the flexible billboard model, which states that binding site position and orientation can be variable for a given enhancer output. Image (c) represents the TF collective model, which suggests that TF-mediated recruitment and TF-TF cooperativity drive enhancer activity, rather than a strict vs. flexible motif grammar paradigm. Adapted from Spitz & Furlong, 2012.

These studies underscore a remarkable operational complexity to enhancer function. Importantly, they suggest a "granular" paradigm of enhancer activity – contributive regulatory elements may not act in stereotyped ways to drive gene expression, but can employ divergent strategies of TF recruitment and multiplexed regulation to enable enhancer activity.

Analysis of Current Methods of *Cis*-Regulatory Site Identification

This section discusses the "imperfect" specificity of TFs to their target genomic sites in relation to enhancer function, the shortcomings of current high-throughput enhancer

definition methodologies, and the applications of the proposed CRISPR/Cas experimental system of functional CRE detection.

Enhancer Prediction from Motif Recognition Analysis

TFs recognize short, 6-12 bp degenerate DNA sequences and exhibit preferential binding to a selection of target genomic sites, whose sequence patterns can be formalized as a TF binding motif (Boeva, 2016). Interestingly, TFs can have variable affinities for the sequences that comprise its binding motif, and certain TFs also can bind relatively dissimilar sequences with the same specificity (Zhao, Ruan, Pandey, & Stormo, 2012). Several mathematical models have been developed in order to represent a TF binding motif, and take into account the disparate set of sequences that a TF can favor. One of the most prevalent models of TF specificity is the position weight matrix (PWM), which specifies the position-dependent probability of each nucleotide within a motif (Boeva, 2016). PWM construction relies on using experimentally determined genomic sequences from ChIP-seq data as the collection of "input sequences" that are examples of a TF binding motif; then, this large input set is condensed into a single sequence logo that prioritizes the nucleotides that had the highest occurrence frequency in a given position (Zhou et al., 2015). Genome-wide enhancer prediction methods using motif finding frequently involve scanning large genomic regions or promoter-proximal areas to identify sequences that have a high alignment score with a known PWM, and extrapolating *cis*-regulatory activity from the presence of multiple motif instances. These approaches are based on a simplifying assumption that high-affinity (or high "scoring") sites are causal to enhancer function in regulating gene expression.

However, recent experiments involving the construction and manipulation of *cis*-regulatory structure have shown that enhancers containing low-affinity binding sites can still effectively mediate gene expression patterns due to optimal binding site grammar (Farley, Olson, Zhang, Rokhsar, & Levine, 2016). As such, current enhancer prediction methods based on motif analysis generalize functional enhancers as "hubs" of high-affinity binding sites, and underappreciate the functional significance of non-canonical regulatory elements as well as the role of a "binding site neighborhood" in promoting enhancer output.

High-throughput Experimental Methods to Identify Enhancer Elements

Genome-wide experiments that chronicle TF occupancy patterns generate a TF binding atlas from which active regulatory regions are inferred, making these approaches susceptible to "transcriptional noise" in the form of non-specific TF binding. In order to effectuate gene expression, TFs need to navigate through a crowded nuclear environment, and sift through the genome to identify target DNA sites for docking (Schmidt, Sewitz, Andrews & Lipkow, 2014). Single molecule tracking experiments have recently demonstrated that throughout this search process, as TFs interpret the genomic information to home in on a target sequence within a *cis*-element, they also non-specifically collide with DNA (Chen et al., 2014). Importantly, studies on the kinetics of TF target searching have indicated that these non-specific transactions with DNA occur frequently and stochastically (Chen et al., 2014).

One of the most employed methods of TF-binding site mapping and putative *cis*-regulatory characterization is ChIP-seq, a technique which involves isolating crosslinked DNA-protein fragments using a TF-specific antibody and high-throughput sequencing of the DNA fragments that were directly bound by the TF (Boeva, 2016). The output of a ChIP-seq experiment is thousands of binding peaks throughout the genome, which are clustered to demarcate CRE domains and frequently coined as "active regulatory regions". In conjunction with recent conclusions on the nature of TF search dynamics within the cell, this suggests that many TF binding sites are not relevant to gene regulation, and only a selective portion of ChIP-derived genomic sites exert a functional influence on target gene expression. Thus, by permissively incorporating non-specific TF-DNA interactions, ChIP-based methods of TF occupancy profiling can yield a high false positive rate of regulatory element discovery. Interestingly, these methods have an additional drawback of being blind to TF-TF cooperative interactions that can potentiate a regulatory element. ChIP experiments do not robustly capture indirect TF-DNA associations that occur by recruitment of a TF or co-factor protein by an enhancer-bound TF (Spitz & Furlong, 2012).

Proposing a New Method to Overcome Current Experimental Limitations

Recent advances in *cis*-regulatory annotation methods include epigenome profiling for enhancer-associated chromatin features. These assays involve surveying large non-coding genomic spaces for nucleosome-depleted regions and specific histone modification marks that are associated with TF occupancy. Frequently, chromatin accessibility is indicative of cooperative TF activity, as certain types of pioneer TFs are

capable of actively repositioning nucleosomes to “open” the DNA for subsequent TF binding (Calo & Wysocka, 2013). Through the described collection of genome-wide tools for enhancer identification, over 400,000 putative enhancers have been identified in the human genome (Calo & Wysocka, 2013).

However, these *de facto* approaches to enhancer classification all use correlative, indirect measures of enhancer activity to impute *cis*-regulatory function on gene expression, and do not match identified enhancers to a target gene. Consequently, existing methods do not enable quantification of the functional significance of candidate enhancers, as they broadly capture non-specific, transient binding to regulatory elements and neither delineate which genomic sites can natively affect transcriptional response nor score enhancers based on the strength of downstream expression modulation. We address these shortcomings by formulating a functional assay for CRE identification that directly surveys and scores enhancer activity in a high-throughput format. The CRISPR/Cas genome editing platform intercalates targeted mutations in >150 kb of genomic space surrounding the *BRCA2* gene, enabling unbiased articulation of the entire complement of regulatory sites that are necessary for *BRCA2* output. The final output is a description of the diverse collection of regulatory elements whose disruption causes a measurable alteration in *BRCA2* gene expression, encompassing the repertoire of sites that hosted binding events that directly and indirectly facilitated stable gene expression from a distal or proximal location.

The Biological Role of BRCA2 in DNA Repair and Genomic Stability

The lack of functional annotations in the non-coding genome has precluded understanding of how variation in focal regulatory nodes can lead to aberrant expression of disease-associated genes. It is challenging to attribute phenotypic significance to non-coding mutations because it is unclear which mutations perturb functional regulatory sequences (Ward & Kellis, 2012). We implement our system of "functional cartography" to characterize *BRCA2*, a DNA repair gene that participates in the homologous recombination (HR) pathway.

The BRCA2 protein mediates repair of double stranded DNA breaks (DSBs) by controlling the localization and recombinase activity of RAD51, a repair protein that catalyzes homologous pairing of a broken DNA strand with its intact sister chromatid template (Gudmundsdottir & Ashworth, 2006). During this process of HR by gene conversion, BRCA2 binds RAD51 and facilitates its transport into the nucleus and to specific sites of DNA damage, where it then assists in RAD51-loading onto a broken 3' overhang of single stranded DNA; this "stabilizing" role of BRCA2 is necessary to promote strand invasion of the broken ssDNA-RAD51 complex into its homologous sister chromatid for high-fidelity DNA synthesis (Prakash, Zhang, Feng, & Jasin, 2015). Deficiency of functional BRCA2 caused by deleterious mutations incapacitates RAD51 mobilization to DSB sites, resulting in inefficient HR repair and genomic instability. As a consequence of HR impairment, cells resort to alternative error-prone mechanisms to repair DNA lesions, resulting in the runaway accumulation of DNA replication errors and mutagenic events from inaccurate repair processes. As such, *BRCA2* is considered a tumor suppressor gene, because the presence of inactivating *BRCA2* mutations

predisposes a “hypermutation” cellular state, contributing to the development of hereditary breast and ovarian cancers. Germline mutations in the *BRCA2* gene account for the majority of familial cases of breast and ovarian cancers, implicating *BRCA2* as a genetic driver of tissue-specific cancer progression (Gudmundsdottir & Ashworth, 2006).

Interpreting *Cis*-Regulatory Variants in Hereditary Cancer Screening

Over 1000 *BRCA2* sequence variants that confer hereditary susceptibility to breast and ovarian cancers have been identified by genetic screening (Maia et al., 2012). To date, mutational screening of *BRCA2* from patient-derived tissue and blood samples involves exon and intron-exon junction coverage, enabling the detection of genetic abnormalities in coding regions and splice sites. Deleterious mutations have been identified throughout the coding framework of the *BRCA2* gene, frequently involving truncating mutations, as well as nucleotide substitutions that disrupt critical protein binding domains (Prakash et al., 2015). However, mutation analysis techniques (nucleotide sequencing, genotyping) exclude the surrounding regulatory genomic space, bypassing any non-coding mutagenic occurrences that predispose the progression of breast and ovarian cancers. Thus, we develop a high-resolution CRISPR/Cas assay that parses through the regulatory genome to precisely map functional regulatory elements that control *BRCA2* expression. The novel system identifies functional regulatory sequences that are causally linked to extinguishing *BRCA2* expression, thus associating these sites with breast cancer risk and enabling clinical prioritization of genetic variation that occurs in these *cis*-regulatory sites.

The development of high-throughput functional mapping techniques can be harnessed to improve the accuracy of genetic risk assessment and disease diagnostics. Currently, genetic screening of hereditary cancer-associated genes (*BRCA2*, *BRCA1*, *ATM*, *RAD51C*) is limited to exon coverage, because it is challenging to attribute a gene-specific regulatory role to *cis*-acting genomic sequences (Walsh, 2015). Functional regulatory assays enable tractable interrogation of a variety of DNA repair genes that are linked to hereditary breast and ovarian cancer susceptibility. By advancing systematic annotation of functional non-coding elements, it is possible to expand the "search space" of current mutational screens to include functionally significant non-coding regions that directly regulate transcriptional output of hereditary cancer genes. This enables identification of deleterious *cis*-regulatory variations in individuals for personalized genomic assessments with enhanced predictive capacity for hereditary breast and ovarian cancer risk.

Chapter II

Materials and Methods

The first section of the Materials and Methods chapter describes the experimental procedures for *BRCA2* gRNA integration into a customized *BRCA2* reporter mouse cell line, fluorescence-based positive selection of expression loss phenotypes, and preparation of cellular genomic DNA for targeted next-generation sequencing. The second section details the computational methods for gRNA library design and analysis of gRNA representation and enrichment levels in GFP^{neg} and GFP^{med} populations.

Experimental Cell Culture Methods

Cell Culture Conditions

Experiments were performed with 129P2/OlaHsd mouse embryonic stem cells (mESCs). mESCs were maintained on gelatin-coated plates feeder-free in mESC media composed of Knockout DMEM (Life Technologies) supplemented with 15% defined fetal bovine serum (FBS) (HyClone), 0.1mM nonessential amino acids (NEAA) (Life Technologies), Glutamax (GM) (Life Technologies), 0.55 mM 2-mercaptoethanol (b-ME) (Sigma), 1X ESGRO LIF (Millipore), 5 nM GSK-3 inhibitor XV and 500 nM UO126. Cells were regularly tested for mycoplasma.

Derivation of the ROSA26 locus gRNA cassette

The purpose of engineering a unique mESC cell line with a knock-in gRNA cassette in the chromatin accessible ROSA26 locus is to enable site-specific incorporation and expression of a singular library gRNA per cell. The gRNA cassette consists of a U6 promoter upstream of a placeholder non-targeting "dummy" gRNA sequence and the gRNA hairpin scaffold, constructed such that the dummy gRNA protospacer will be subsequently replaced by a genome-targeting gRNA upon *BRCA2* gRNA library electroporation.

To successfully knock-in a dummy gRNA expression cassette, it is first necessary to PCR a plasmid containing the U6 promoter and dummy gRNA plus hairpin scaffold with primers that amplify the entire expression construct and attach ROSA26 homology arms (Table 3). The PCR reaction is performed as a 35 cycle 2-step PCR (98 for 10 seconds, 72 for 45 seconds), resulting in a 750 bp fragment. The resulting purified PCR product is subjected to a sequential PCR (35 cycle 2-step PCR, 98 for 10 seconds, 72 for 45 seconds) to extend the ROSA26 homology arms, resulting in an 800 bp final amplicon (Table 3). Upon completion of the PCR amplification steps, mESCs at a cell density ~ 1.0e6 are co-electroporated with 5 ug p2T CBh *S.Pyogenes* Cas9 BlastR, 5 ug p2T U6sgROSA26-FE HygroR, and the purified ROSA-HDR dummy gRNA cassette amplicon. Electroporation is performed using a Bio-Rad electroporator set to 230 V, 0.500 uF and maximum resistance. Cells are transiently selected with Blasticidin and Hygromycin for 24-72 hours post-electroporation. Knock-in positive clones are identified by genomic DNA PCR testing of the cassette sequence and sequence verification of the full ~900 bp region, resulting in positive identification of a heterozygous ROSA26 gRNA

expression cassette knock-in cell line (Table 3). The final sequence of the ROSA26 dummy gRNA expression module is:

TCCCATTTTCCTTATTTGCCCCTATTAAAAAACTTCCCGACAAAACCGAAAAT
CTGTGGGAAGTCTTGTCCCTCCAATTTTACACCTGTTCAATTCCCCTGCAGGA
CAACGCCCACACACCAGGTTAGCCTTTAAGCCTGCCCAGAAGACTCCCGCCC
AGCATGTGAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGATACA
AGGCTGTTAGAGAGATAATTGGAATTAATTTGACTGTAAACACAAAGATATT
AGTACAAAATACGTGACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTT
TAAAATTATGTTTTAAAATGGACTATCATATGCTTACCGTAACTTGAAAGTAT
TTCGATTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGAGGCGTC
TGGGTGGCTCTTGGTTTAAGAGCTATGCTGGAAACAGCATAGCAAGTTTAAA
TAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTT
GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTTTTAGCGC
GTGCGCCAATTCTGCAGACAAATGGCTCTAGAGGTACGGCCGCTTCGAGCAG
ACATGATAAGATACATTGATGAGTTTGGACAAACCACAACTAGAATGCAGTG
AAAAAAATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCA
TTATAAGCTGCAATAAACAAGTTAACAACAACAATTGCATTCATTTTATGTTT
CAGGTTCAGGGGGAGATGTGGGAGGTTTTTTAAAGCAAGTAAAACCTCTACA
AATGTGGTAAAATCGCGATGCAGATCACGAGGGAAGAGGGGGAAGGGATTC
TCCCAGGCCCAGGGCGGTCCTCAGAAGCCAGGAGGCAGCAGAGAACTCCCA
GAAAGGTATTGCAACACTCCCCTCCCCCCTCCGGAGAAGGGTGCGGCCTTCT
CCCCGCCTACTCCAC

The dummy gRNA protospacer sequence is indicated in red.

Generation of the *BRCA2-GFP* Reporter Cell Line

The gRNA screening process is dependent on the generation of a locus-specific GFP knock-in reporter cell line that labels the *BRCA2* gene with a GFP tag for rapid fluorescent signal readout of gene expression and fluorescence-based cell sorting. GFP transgene knock-in is performed using mESCs with an integrated ROSA26 gRNA cassette, as described in a previous protocol. To construct the *BRCA2-GFP* fusion gene, a *BRCA2* exon 27-targeting gRNA is cloned into a plasmid containing a U6 promoter, gRNA hairpin scaffold, and Hygromycin resistance cassette. The *BRCA2*-targeting

gRNA specifies the precise location of CRISPR/Cas-mediated cleavage and GFP transgene insertion into the genome. Next, a GFP insertion cassette is synthesized by two successive PCR amplification steps of the GFP frame, with homology arm primers that add 70-80 bp of *BRCA2* homologous sequence surrounding the desired insertion site to facilitate GFP knock-in via homologous recombination (Appendix, Table 3). Engineered mESCs with the ROSA26 dummy gRNA construct are then co-electroporated with the *BRCA2*-targeting gRNA plasmid, the homology arm-extended GFP amplicon, and a *S.Pyogenes* Cas9 plasmid with a Blasticidin resistance cassette (Figure 2).

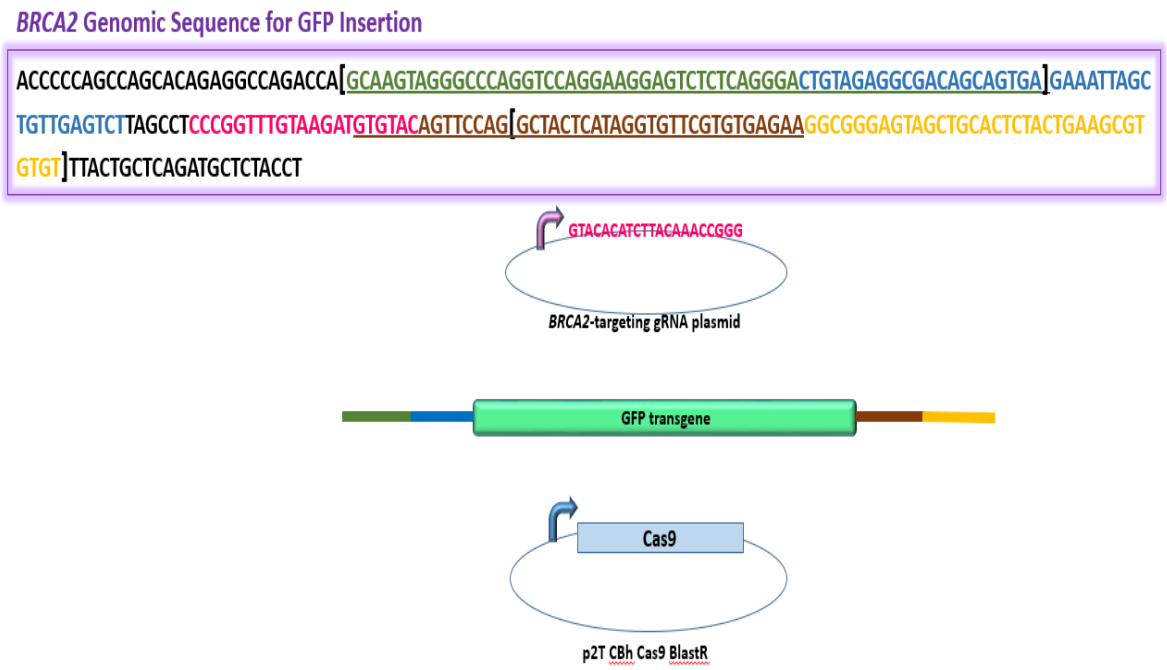

Figure 2. Construction of the *BRCA2-GFP* mESC line. Exon 27 of the *BRCA2* gene is targeted by a site-specific gRNA plasmid, Cas9 plasmid, and GFP transgene amplicon for CRISPR-mediated cleavage and GFP knock-in.

Following Blasticidin and Hygromycin antibiotic selection, the electroporated mESCs are flow cytometrically sorted and GFP-expressing single cells are collected and clonally expanded. Genomic DNA PCRs of clones confirm the locus-specific integration of the GFP tag (Table 3). Flow cytometry analysis of a sequence-verified *BRCA2-GFP* fusion mESC population derived from expansion of a single positive colony reveals a pattern of stochastic *BRCA2* gene expression across a cellular population. Despite multiple rounds of fluorescence-based purification, the *BRCA2-GFP* mESC population is persistently heterogeneous with ~80% of cells as GFP-expressing and ~20% as GFP-negative at a given timepoint, suggesting that cells can dynamically fluctuate between strong and subdued *BRCA2* expression states (Figure 3).

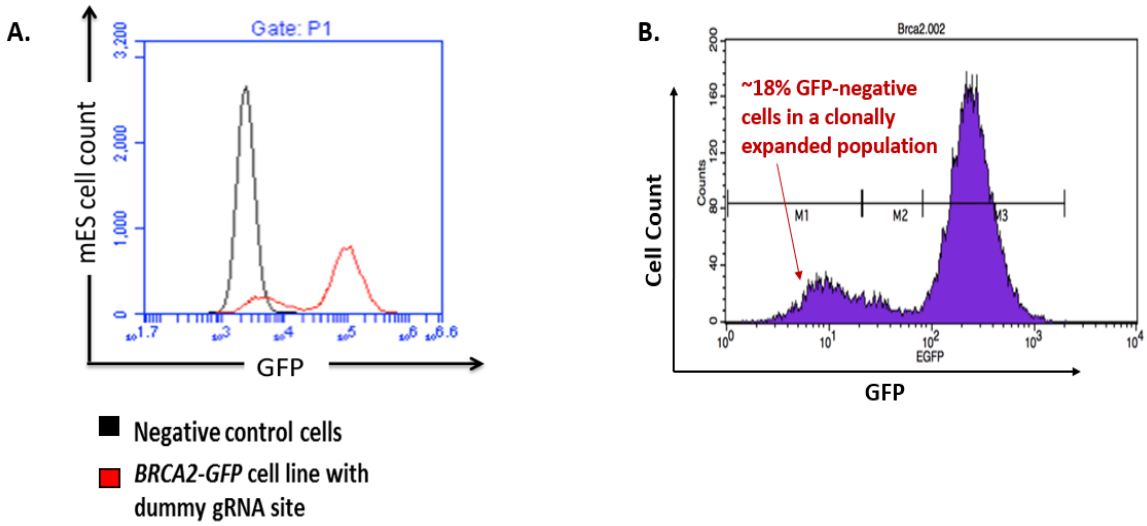

Figure 3. Flow cytometry analysis of the original *BRCA2-GFP* cell population. Figure (3a) compares the *BRCA2-GFP* cell line to a negative control, non-fluorescent cell line to demonstrate GFP fluorescence in the constructed cells. Figure (3b) is a univariate histogram that displays cell count vs. relative fluorescence. The flow cytometry histogram of *BRCA2-GFP* cells reveals a bimodal population distribution, with ~18% of cells landing below the GFP-positive gate.

BRCA2 gRNA Library Screening Process

The *BRCA2* gRNA library comprehensively spans 185 kb of the regulatory genome to uncover functional *cis*-elements that are necessary for *BRCA2* expression in a native context. The assay harnesses the CRISPR/Cas platform to tile mutations across the *BRCA2* regulatory landscape, selects for cells that undergo complete or partial expression loss, and cultivates 3 phenotypically-distinct cellular populations that can be compared to identify gRNAs preferentially associated with *BRCA2* expression loss.

Preparation and Introduction of gRNA Pool into Customized mESCs

The pooled *BRCA2* gRNA library is PCR amplified to attach ~80-90 bp of ROSA26 homology arms to each side (5' and 3') of the gRNAs (Table 3). The PCR is performed as an 800uL NEBNext reaction for 35 cycles (3 step PCR; 98 for 10s, 62 for 30s, 72 for 30s) using 1% of the gRNA library, yielding ~189 bp gRNA amplicons. *BRCA2-GFP* fusion mESCs with the ROSA26-integrated gRNA cassette are co-electroporated with 80 ug p2T CBh Cas9 BlastR, 80 ug of a gRNA plasmid that cleaves the dummy gRNA protospacer, and the purified ROSA26-extended gRNA library. The sequence of the gRNA that cuts the dummy gRNA is GAAACACCGAGGCGTCTGGG.

At the time of electroporation, the mESCs have grown to 80% confluence on 2 15 cm plates – a starting density of ~2e7 cells/plate is requisite for >1e7 mESCs to survive antibiotic selection to preserve library diversity in the cellular population. 24 hours post-electroporation, the cells are subjected to Blasticidin treatment (1:1000) for the following

48 hours.5 days post-electroporation (d5), ½ of the surviving mESCs are pooled on the same 15 cm tissue culture plate, and the remaining ½ is frozen and stored at -80.

Flow Cytometry Separation of GFP^{neg} and GFP^{med} Populations

Following *BRCA2* library gRNA integration in the ROSA26 locus and gRNA-induced mutagenesis at a complementary genomic site, the targeted cell population is a mixed pool of GFP-positive, GFP-negative and GFP-medium phenotypes. On day 7 post-electroporation (d7), genomic DNA from ¼ of the merged bulk population is collected, and ½ of the mESC bulk population is flow cytometrically sorted according to GFP expression loss. In the first round, a positive vs. negative flow cytometry gate permissively segregates high and low GFP intensity, and captures single cells with both intermediate and complete fluorescence loss; in this first round of purifying selection, ~23-24% of the total sorted population is encompassed. The sorted GFP^{neg} and GFP^{med} cells are subsequently cultured for 3-4 days, and on d10-d11, the cells are subjected to a second fluorescence sorting. In the second round of sorting, the cells are partitioned into a GFP^{neg} or GFP^{med} population based on the extent of GFP expression, and the two separated populations are cultured for 2-3 days. Between d12-14, a third fluorescence sorting is performed if necessary to obtain a purified population, based on the noisiness of separation in previous sorting rounds. Following 2 sorts, the GFP^{neg} population is sufficiently purified at >90%, and genomic DNA is collected from ½ of the GFP^{neg} cell population between d12-14. The GFP^{med} population is noisier, with infiltrating positive and negative events, so GFP^{med} mESCs are sorted for a third time, cultured for 2-3 days post-sorting, and genomic DNA is harvested from ½ of the population (Figure 4).

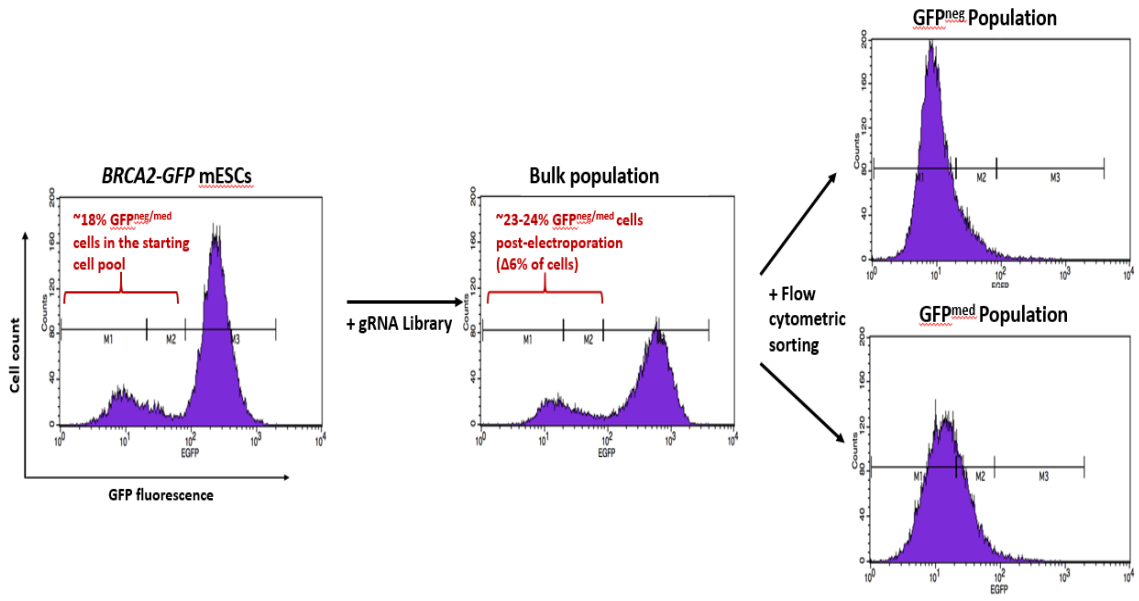

Figure 4. Multi-stage flow cytometry analysis through library targeting and cell sorting. Successive flow cytometry analysis following library introduction and fluorescence sorting reveals a population shift in GFP expression and successful purification of $GFP^{negative}$ and GFP^{medium} sub-populations.

Library Preparation for Genomic DNA Sequencing

Genomic DNA extracted from the three populations (bulk, GFP^{neg}, GFP^{med}) is prepared for Illumina sequencing by a 3-stage PCR process. The first PCR reaction exclusively amplifies integrated library gRNAs in the ROSA26 locus to enable out-competition of the ROSA26-incorporated gRNAs over unincorporated gRNA homology fragments. The first library prep PCR is performed as a 15 cycle NEBNext reaction, with a ratio of 16 ug genomic DNA in an 800 uL volume and a primer concentration of 500 nM (Table 3). This ratio of ~20 ng genomic DNA/uL of reaction volume is necessary to avoid over-saturation of DNA template for amplification. The purified PCR product is then tested in a 20 uL SybrGreen qPCR reaction using 0.1 uL of the PCR product (Table

3). The qPCR count typically falls between 9-14 cycle counts, and the qPCR count divided by 2 is then used as the cycle value for the second library prep PCR.

The second PCR barcodes the samples with a 5 bp sample-specific sequence for multiplexing the three genomic libraries. The second PCR is done in a 50 uL NEBNext reaction using 23 uL of the purified product from the first PCR and a cycle number determined from the qPCR value. Primers for the second PCR are added at a 500 nM concentration, and include a forward Illumina PE1 barcode sequence and the reverse library gRNA PE2 primer (Table 3). From this reaction, a sample barcode is introduced between the library gRNA and the PE1 primer, and the first half of the PE2 sequence is attached to the constructs. The purified second PCR product is tested in a 20 uL SybrGreen qPCR reaction using 0.1 uL of the purified product, and the equivalent qPCR count value is applied as the number of cycles for the third library prep PCR.

The third PCR adds the rest of the standard Illumina paired-end sequencing adaptors to the library gRNA fragments. The PCR step is performed in a 50 uL NEBNext reaction using 23 uL of the purified second PCR product and a cycle number determined from the qPCR count. The Illumina PE1 and PE2 primers are included at a 500 nM concentration (Table 3). The final amplicons are ~211 bps, and are composed of a genome-integrated library gRNA sequence, a sample barcode shared across the population, and flanking Illumina paired-end sequencing primers. The final amplicons are purified, quantified for fragment size, and sequenced using an Illumina MiSeq instrument.

BRCA2 gRNA Library Design

The *BRCA2* gRNA library is composed of 10376 gRNAs that collectively target a -93 kb to +93 kb genomic expanse surrounding the *BRCA2* promoter, 74 positive control gRNAs that directly mutagenize the GFP open reading frame (ORF), and 126 negative control non-targeting gRNAs that do not possess genome complementarity. The key design specification is that the gRNA library tiles the *cis*-regulatory genome in an unbiased manner such that known regions are not prioritized over uncharacterized sites and there exists commensurate likelihood of functional element discovery across different genomic categories.

gRNAs were designed using the following algorithm:

1. Establish a broad window of interest for examining the genome to identify regulatory sites of significance (~185 kb of genomic landscape).
2. Find all NGG occurrences on both the forward and reverse strand. The *S.Pyogenes* Cas9 nuclease recognizes the trinucleotide protospacer adjacent motif (PAM) in the genomic DNA, which is necessary for gRNA-DNA pairing and cleavage of the target site.
3. Design the gRNA such that it has 19-20 bps of homology immediately preceding the genomic 5'-NGG-3' PAM.
    a. If the genome sequence is GNNNNNNNNNNNNNNNNNN NGG ($GN_{19}NGG$), the guide RNA spacer sequence (the 5' gRNA targeting

sequence with ~20 bp complementarity with the desired genomic DNA) should be GNNNNNNNNNNNNNNNNNN (GN_{19}).

b. If statement a is not satisfied but GNNNNNNNNNNNNNNNNNN NGG ($GN_{18}NGG$) is satisfied, the gRNA spacer sequence should be GNNNNNNNNNNNNNNNNN (GN_{18}).

c. If statements a and b are not satisfied, the gRNA spacer sequence should be GNNNNNNNNNNNNNNNNNNN (GN_{20}) where the genomic sequence is NNNNNNNNNNNNNNNNNNN NGG ($N_{20}NGG$) – it does not matter if the first G is in not the genome.

For all gRNAs, the presence of the 5'G at the start of the spacer sequence improves U6 transcription.

4. Each designed gRNA spacer sequence is placed in the following template, which is 98-100 bp long:

TTATATATCTTGTGGAAAGGACGAAACACC[$GN_{18/19/20}$]GTTTAAGAGCT ATGCTGGAAACAGCATAGCAAGTTTAAATAAGGCTAGT.

The template consists of a gRNA-proximal stretch of the U6 promoter, the genome-targeting gRNA spacer, and a partial component of the gRNA scaffold. PCR-based extension of the gRNA template completes the U6 promoter and gRNA scaffold sequences using ROSA26 homology arm primers prior to electroporation of the *BRCA2* gRNA library.

Data Processing and Analysis of gRNA Significance

Mapping of MiSeq Reads

The expected full sequence consisting of the sample barcode, primers and designed gRNA is compared to the output reads. Counts for each gRNA for either GFP^{neg}, GFP^{med} or bulk populations are obtained by counting the number of sequenced reads that show exact matches to the designed gRNA.

Visualization of gRNA Read Distributions and Genome Feature Boundaries

The UCSC genome browser (mouse GRCm38/mm10 assembly) is used to visualize the data and create genomic view snapshots for regulatory regions of *BRCA2*. Absolute read counts of gRNAs in the GFP^{neg} and GFP^{med} populations are plotted in the browser, along with a track representing the genome coverage of bulk reads.

Enhancer predictions are made using 6 histone modifications from ENCODE data trained on p300 binding site data from mouse embryonic stem cells. Enhancers are separated into "strong" and "weak" descriptors based on presence of H3K27ac at levels greater than input. Enhancer boundaries are further clarified using established edge-detection methods (Rajagopal et al, 2016). Similarly, DNaseI hypersensitivity hotspots are identified with a standard algorithm utilized by Rajagopal et al. H3K4me3 and H3K27ac signals are displayed in the UCSC browser panel, which are histone

modification marks associated with promoters and active regulatory elements, respectively.

Detection of Significantly Enriched gRNAs in GFP^{neg} and GFP^{med} Populations

The enrichment score of each gRNA in the GFP^{neg} and GFP^{med} populations represents the log fold-change between the sorted and bulk reads, calculated as the $\log_2$ ratio of GFP^{neg} to bulk reads and GFP^{med} to bulk reads. The term "enriched" signifies a $\log_2$ ratio of GFP^{sorted}/bulk > 0. Statistically significant *BRCA2*-targeting gRNAs in GFP^{neg} and GFP^{med} populations are discerned by the following condition:

For a given sorted population (GFP^{neg} or GFP^{med}), if the $\log_2$ ratio of GFP^{sorted}/bulk for targeting $gRNA_i$ > 2 standard deviations above the mean $\log_2$ ratio of GFP^{sorted}/bulk of all negative control gRNAs, then $gRNA_i$ is considered significantly enriched in that population.

Detection of Significant gRNA Windows in GFP^{neg} and GFP^{med} Populations

A "sliding window" approach is employed to identify the presence of significant regions across N consecutive gRNAs, where N = 25. The "sliding window" approach tests whether the mean GFP^{neg} read count of the 25 gRNAs in each window is significantly greater than the mean GFP^{neg} read count of the 126 negative control gRNAs using a statistical t-test (Fulco et al., 2016). The Benjamin-Hochberg procedure is applied

to compute adjusted p-values, controlling the false discovery rate (FDR) at level $\alpha = 0.05$ across multiple hypothesis comparisons.

Chapter III

Results

Here we present a high-resolution CRISPR/Cas-based assay that quantifies the functional impact of the regulatory genome in its native cellular context. The assay utilizes a uniquely designed ~10,000 sequence gRNA library to comprehensively screen >150 kb of genomic space in an unbiased manner to uncover the distribution of functional elements that control *BRCA2* expression. Genomic perturbation is achieved by the nuclease action of Cas9, which co-localizes with a single library gRNA per cell to the targeted genomic regulatory site and induces DNA cleavage, resulting in site-specific mutagenesis. This method utilizes a fluorescence-based readout to obtain accurate measurement of *BRCA2* expression and enable cell population separation based on the degree of expression loss. Next-generation sequencing of filtered populations yields quantitative information on the differential enrichment of gRNAs that target a sequence encoding a regulatory function, and concatenating this information across thousands of tiled gRNAs produces a functional regulatory architecture of all necessary *cis*-elements that power *BRCA2* expression. This technique confers 3 benefits to a standard high-throughput screening process: first, the gRNA library size confers unbiased targeting of a large genomic region; second, it uses a cloning-independent system for library gRNA integration with a single targeting event per cell; third, a fluorescent reporter is used for efficient gene expression measurement (Rajagopal et al., 2016) (Figure 5).

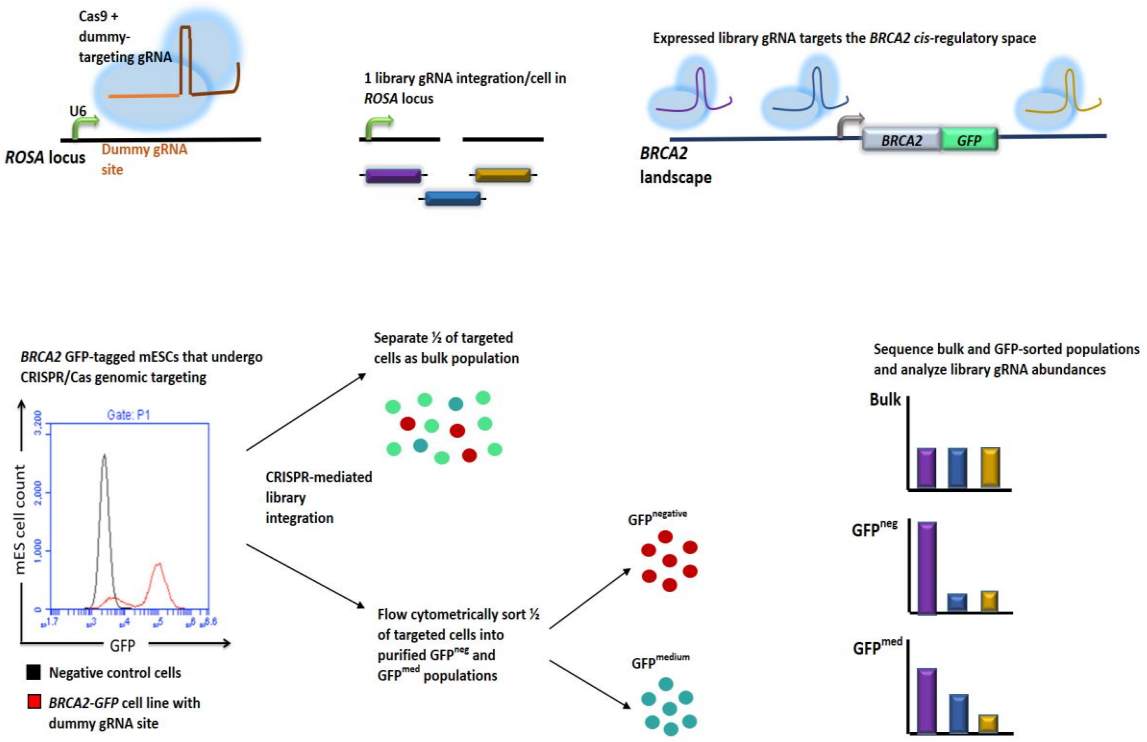

Figure 5. Assay workflow for high-throughput gRNA screening. Library gRNAs are tiled across the *cis*-regulatory regions of the GFP-tagged *BRCA2* locus and cells are sorted according to their extent of expression loss. Deep sequencing identifies the gRNAs that induce complete and partial expression loss.

Assessment of gRNA Library Integration Efficiency and Representation

Efficient HR-mediated integration of library gRNAs into the *ROSA26* locus is necessary for robust library diversity (percentage of different gRNAs detected from the original pool). Out of the 10576 gRNAs in the *BRCA2* library, 10348 gRNAs are detected in the bulk population (>= 1 matched sequencing read). Thus, ~98% of library gRNAs have at least one faithful genomic integration event that is captured by bulk population

sequencing, resulting in strong library diversity and minimal gRNA drop-out during the integration step.

Next, we analyze gRNA abundances in GFP^{neg} and GFP^{med} populations, detecting retention of 4692/10576 gRNAs in the GFP^{neg} population and 5652/10576 gRNAs in the GFP^{med} population. Fluorescence-based purifying selection filters out ~50% of gRNAs that target non-contributory or nonfunctional regulatory sites; however, while targeting of a nonfunctional cis-regulatory node accounts for a majority of the occurrences of gRNA depletion during fluorescence selection of *BRCA2* expression loss, it is possible that a small subset of the unretained gRNAs hit upon sequences with regulatory significance but fail to inactivate their functionality. As expected, a high fraction (70/74, or 95%) of positive control gRNAs was retained in the GFP^{neg} population, as these gRNAs were specifically designed to induce mutations in the coding frame of the GFP reporter. A lower fraction of negative control gRNAs (71/126) was represented with at least 1 sequencing read in the GFP^{neg} population, but these gRNAs may need to be further examined to ensure that there is no genome complementarity or off-target cutting that can potentially induce expression loss. Figure 6 depicts scatterplots of log-transformed read counts for each represented gRNA in the GFP^{neg} and GFP^{med} populations.

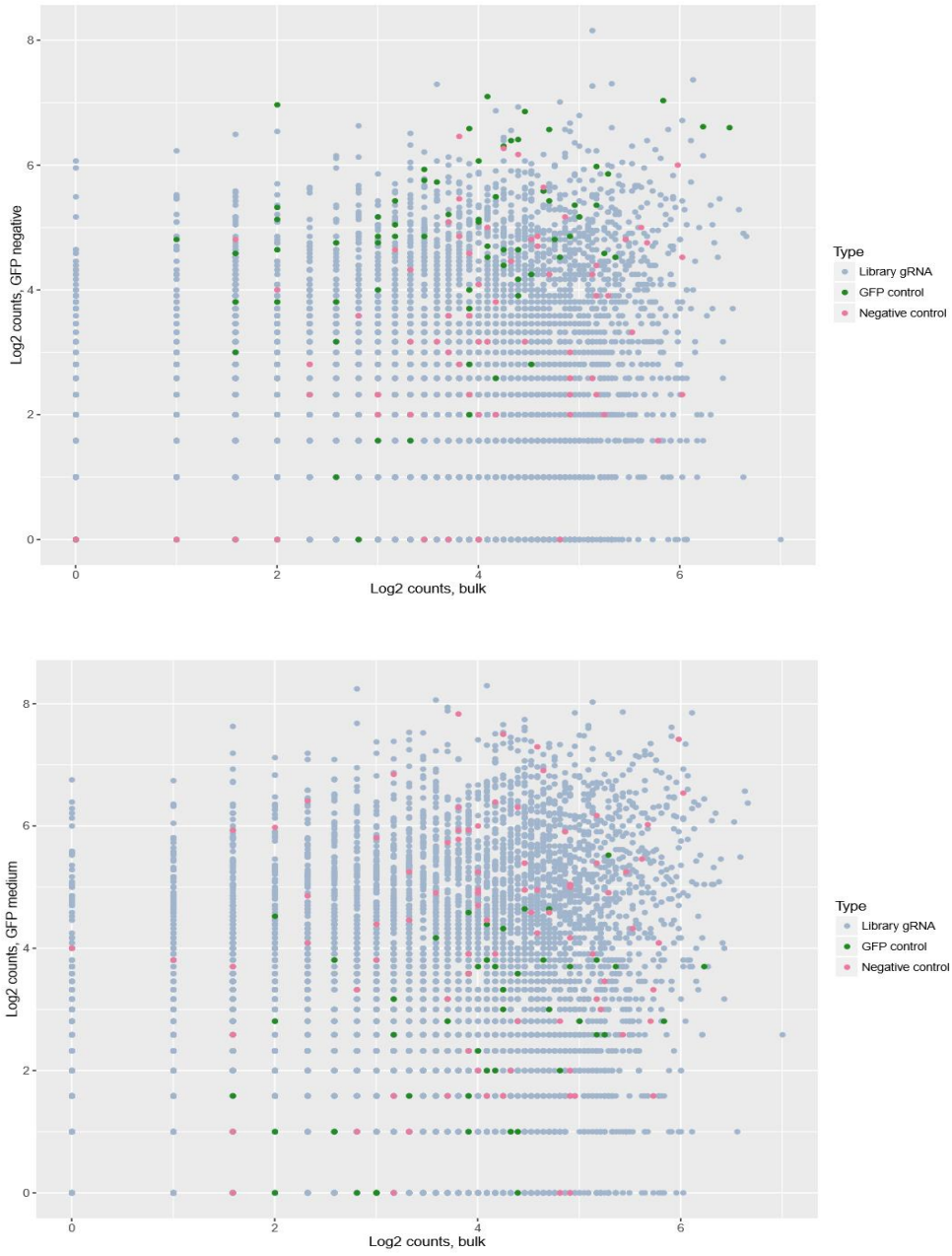

Figure 6. Log_2 sorted counts vs. log_2 bulk counts for each gRNA. The upper figure represents a scatterplot of log_2 GFP^{neg} counts vs. log_2 bulk counts for each gRNA with representation in the GFP^{neg} population. The lower figure depicts a scatterplot of log_2 GFP^{med} counts vs. log_2 bulk counts for each gRNA with representation in the GFP^{med} population.

Genomic Mapping of gRNAs

Here we map the distribution of gRNA read counts in the GFP^{neg} and GFP^{med} population by their genome coordinates to identify 4 key *BRCA2* regulatory trends: first, the marginal abundance ("regulatory strength") and density ("regional importance") between clusters of gRNAs, the representation of gRNAs that coincide with known genomic categories and regulation-associated markers, the representation of gRNAs in regions devoid of typical gene regulatory signatures, and the pattern of genomic feature diversification within the *BRCA2* regulatory architecture.

The assay is designed such that the relative abundance of a given gRNA, calculated as the fold-change ratio of gRNA prevalence between the sorted and bulk populations, is generally correlated to the functional importance of that element to *BRCA2* expression. While extenuating aspects of the CRISPR screening platform – such as the gRNA cleavage efficiency and the likelihood that a gRNA-induced mutation will inactivate the regulatory sequence – affects the relationship between gRNA relative abundance and *cis*-regulatory activity, the assay is designed to mitigate this effect by incorporating numerous gRNAs that target flanking or overlapping sequences, and more broadly, by extensive genome coverage. A genomic visualization of gRNA abundance across the array of spatially-mapped gRNAs elucidates a distributed regulatory topology governing *BRCA2*, whereby individual gRNA peaks (high-abundance gRNAs) are located in distal zones and proximal regions to the *BRCA2* promoter, but no single gRNA window commands the regulatory signal that controls gene expression. Similarly, we

observe that gRNA density is relatively uniform, without discernible formations (large clusters or gaps) of represented gRNAs along the GFP^{neg} and GFP^{med} tracks (Figure 7).

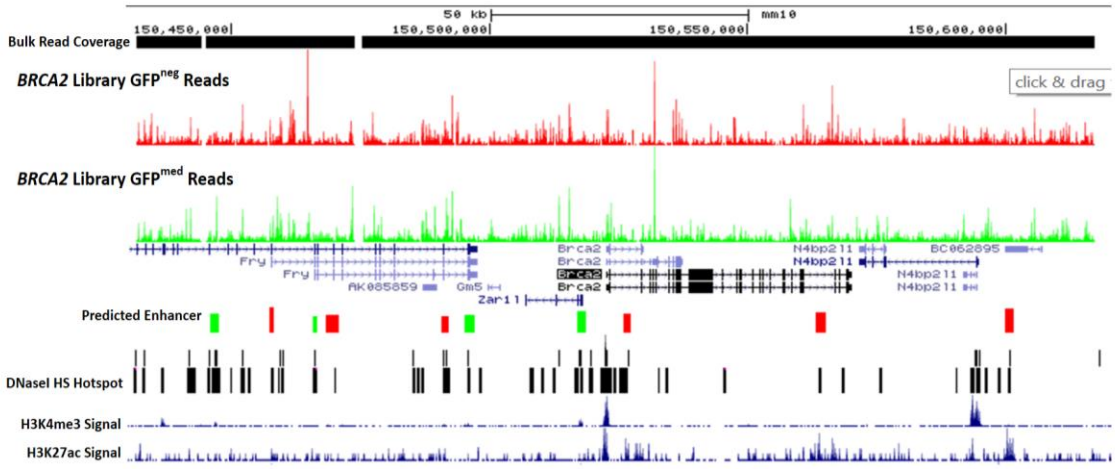

Figure 7. UCSC browser display of the 185 kb genomic space proximal to *BRCA2*. In top to bottom order, the genomic view provides a track of bulk read coverage across the targeted region, spatially mapped GFP^{neg} and GFP^{med} gRNAs with bar height proportional to gRNA abundance in the given population, annotated genes, predicted strong (red) and weak (green) enhancers, DNaseI hotspot regions, and H3K4me3 and H3K27ac ChIP-seq signals.

Next, we focus on the distribution of gRNAs within established categories of genomic elements associated with regulatory potential and unmarked DNA regions. Genomic categories include non-coding RNA transcripts (ncRNAs) residing within gene loci and intergenic stretches, strong and weak enhancers predicted from chromatin modifications, DNaseI hypersensitive hotspots, and histone acetylation and methylation peaks. While these molecular hallmarks are frequently used to infer active regulatory participation in controlling proximal gene expression, causality between these structural

or epigenetic features and the regulatory function of the underlying sequences has not been systematically established (Sanjana et al., 2016). The assay is developed to address this specific caveat by first directly scoring for the contribution of *cis*-acting sites to *BRCA2* expression, and subsequently intersecting the results with ENCODE-derived genomic annotation data to determine the co-occurrence of gRNA representation and enhancer-associated marks. Figure 8 depicts gRNA representation within selected genomic elements, including a strong predicted enhancer within *FRY*, a long intergenic non-coding RNA (lncRNA), the H3K4me3 peak in the *BRCA2* promoter, an unmarked regulatory element (URE) within *BRCA2*, and the H3K4me3 peak in the *N4BP2L1* promoter.

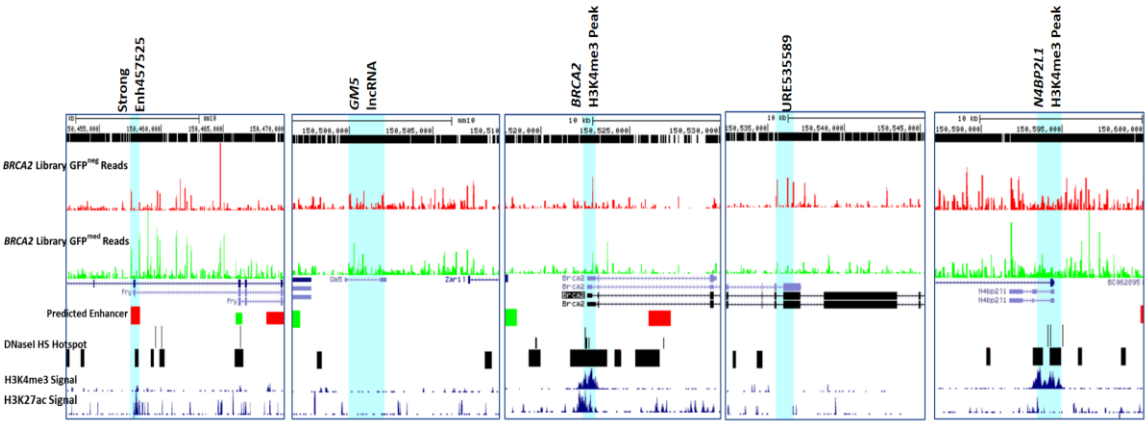

Figure 8. Snapshots of genomic regions with corresponding gRNA abundance plots and UCSC browser annotations. For elements Strong Enh457525 and URE535589, the last 6 digits represent the genomic start coordinates 150XXXXXX. In left to right order, the first slide depicts a strong predicted enhancer element located 65 kb upstream of *BRCA2*, the second depicts the GM5 lncRNA located in an intergenic stretch between *FRY* and *BRCA2*, the third depicts the *BRCA2* H3K4me3 peak that coincides with the gene promoter region, the fourth depicts an unmarked regulatory region, and the fifth depicts the *N4BP2L1* H3K4me3 peak located 72 kb downstream of the *BRCA2* start site.

Through high-resolution mapping of gRNAs in the GFP^{neg} and GFP^{med} populations, we can observe individual gRNAs with maximal abundance within a given genomic element and compare differential densities of represented gRNAs across genomic regions. Interestingly, URE535589 contains several gRNAs with high abundance in the GFP^{neg} population, despite the fact that this region lacks canonical characteristics of regulatory activity, such as H3K27ac and DNaseI hypersensitivity. The H3K4me3 peak region of the *N4BP2L1* gene promoter has a high density of represented gRNAs, suggesting the involvement of specific sites within the *N4BP2L1* promoter in regulating *BRCA2* expression.

We then quantify the fraction of represented gRNAs in GFP loss populations among the different genomic categories to determine the prevalence of regulatory participation from diverse elements. Overall, the fractions of represented gRNAs across elements from various genomic backgrounds reveal that diversified regulatory inputs govern *BRCA2* expression. While there are disparities in the proportions of represented gRNAs between 2 elements from the same genomic category – for instance, URE581139, located downstream of *BRCA2*, has a lower fraction of GFP^{neg}-represented gRNAs than URE535589 – the results support the conception of a distributed repertoire of functional elements regulating *BRCA2*. We also observe differential fractional representation of gRNAs between GFP^{neg} and GFP^{med} populations, suggesting that some targeting gRNAs only induce partial expression loss, either as an artifact of the gRNA mutagenic screening and selection process, or as a product of a complex regulatory code whereby some functional elements tune the rate or efficiency of transcription (Figure 9).

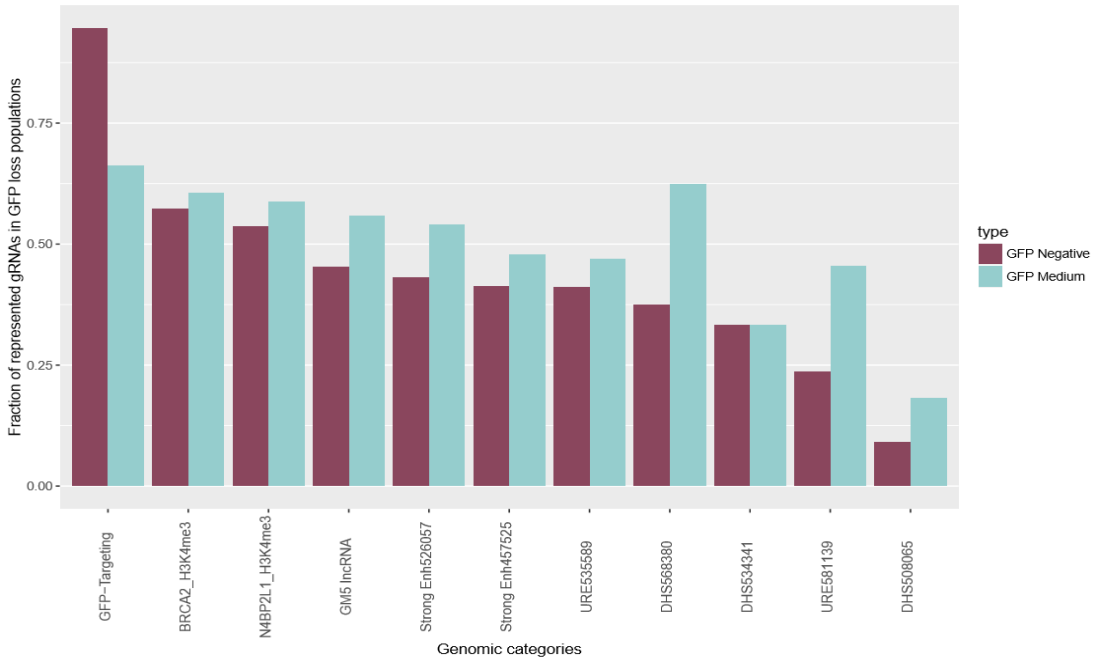

Figure 9. Fraction of GFP^{neg} and GFP^{med} represented gRNAs across different genomic categories. Fractions are calculated as the proportion of targeting gRNAs that map to a given element with >= 1 read count for each respective population. Genomic categories are established classes of elements associated with gene regulation, as well as elements that lack UCSC-designated markers of regulatory activity. The last 6 digits for Strong Enh, URE and DHS elements indicate the genomic start coordinates 150XXXXXX.

Identification of Significant gRNAs by Differential Enrichment

Here we identify gRNAs that are significantly enriched in GFP loss populations by thresholding gRNA enrichment scores, and impute regulatory significance from statistically striking count differences. Enrichment scores are calculated as the log fold-change ratio of gRNA abundance between the sorted and bulk populations (the relative abundance), representing the functional importance of the *cis*-regulatory site. Figure 10 depicts the respective cumulative distribution function of the log fold-change ratio of represented gRNAs in the GFP^{neg} and GFP^{med} populations.

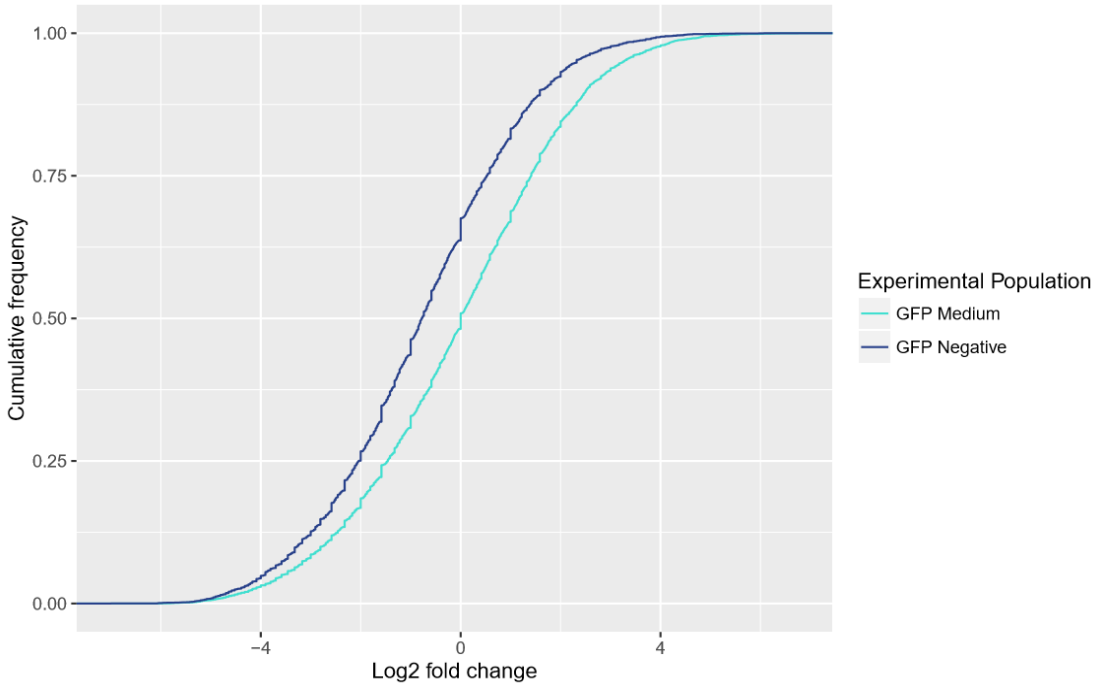

Figure 10. Cumulative distribution function plots of log fold-change ratios. CDF plots show the $\log_2$ fold changes in gRNA abundance between sorted and bulk populations for GFP^{neg} and GFP^{med} represented gRNAs.

Detection of Significantly Enriched gRNAs in GFP^{neg} and GFP^{med} Populations

Around 40% of GFP^{neg} gRNAs and 50% of GFP^{med} gRNAs are enriched with a log fold-change ratio > 0, but marginal enrichment is frequently a by-product of chance in a multi-stage screening process, sequencing-introduced error, or background from stochastic variations in *BRCA2* transcription. As such, we utilize an enrichment score threshold to selectively capture gRNAs with significant over-representation in GFP loss

populations, setting a cutoff of 2 standard deviations above the average relative abundance of the set of negative control gRNAs. This method of gRNA enrichment classification enables high-resolution detection of short genomic sequences (~20 bps) that are necessary for *BRCA2* expression (Figure 11).

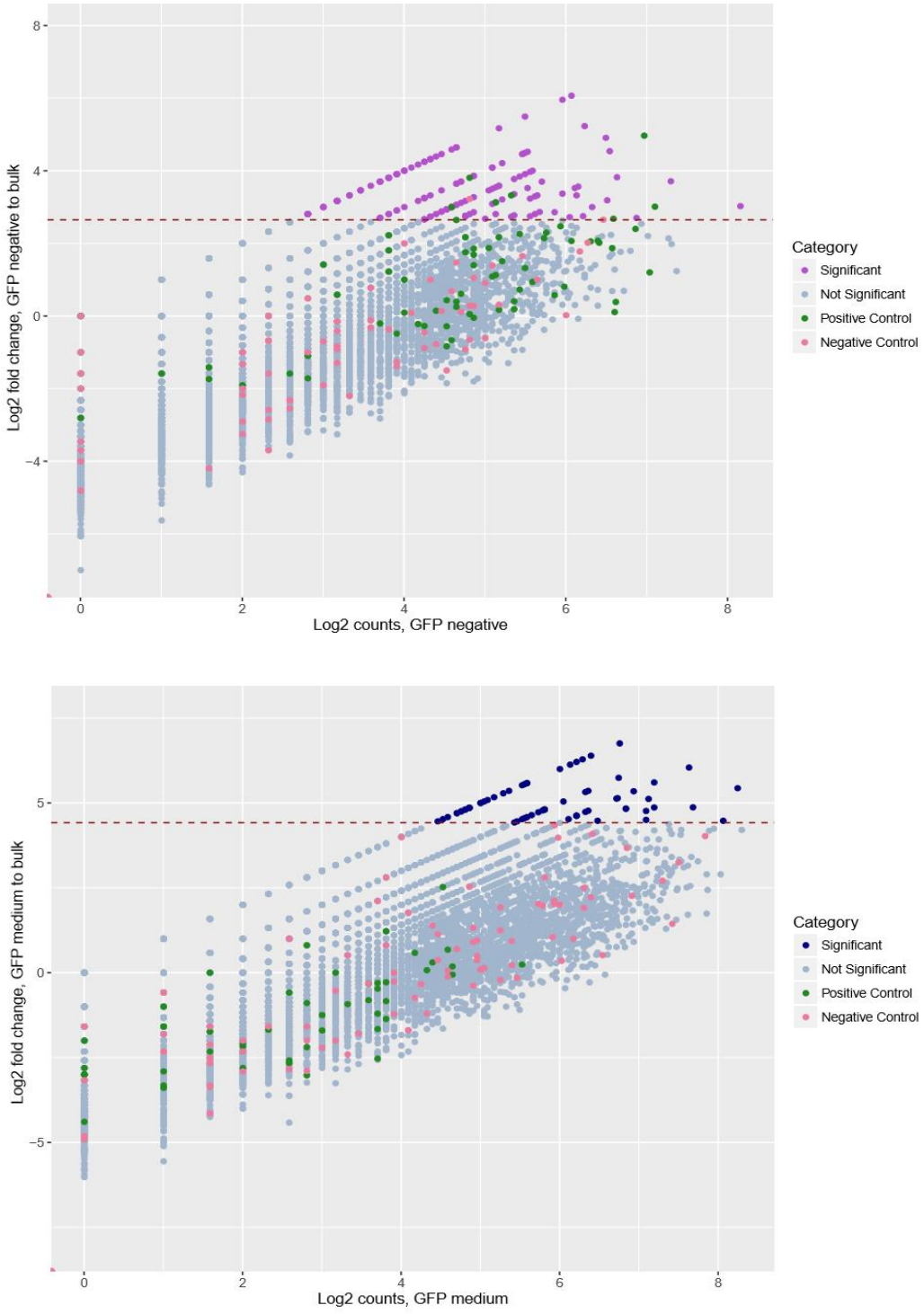

Figure 11. Log_2 fold-changes vs. log_2 sorted counts for GFP^{neg} and GFP^{med} gRNAs. The upper scatterplot depicts the log-transformed GFP^{neg}-to-bulk fold-change values, and the lower scatterplot depicts the log-transformed GFP^{med}-to-bulk fold-change values. In each plot, the dashed line indicates the enrichment score cutoff for significance, and gRNAs above the threshold are demarcated as significantly enriched.

A total of 153 gRNAs are significantly enriched over the set of non-targeting controls in the GFP^{neg} population, and 65 gRNAs are significantly enriched in the GFP^{med} population. There is a considerable overlap of 44 significant gRNAs that are present in both populations. From these results, we conclude 2 things: first, that we are able to use purifying selection and enrichment thresholding to generate a filtered panel of highly enriched gRNAs (~2% of the original 10376 *BRCA2* gRNA targeting library) that are preferentially associated with expression loss; and second, the overlap of significant gRNAs in the GFP^{neg} and GFP^{med} populations reinforces that targeted disruption of these sites affects transcriptional output, but the extent of transcriptional alteration (partial or complete) for a shared gRNA may depend on the type of induced mutation at that element (Table 1).

Table 1.

Top-enriched gRNAs in GFP^{neg} and GFP^{med} cells.

Top-hit GFP^{neg} gRNAs	Genome Location	Log fold-change of sorted to bulk
Top-hit #1	Intron 24/26 in BRCA2	6.06
Top-hit #2	Intron 50/60 in FRY gene	5.95
Top-hit #3	AK085859 lncRNA (within intron 58/60 of FRY)	5.49

Top-hit GFP^{med} gRNAs	Genome Location	Log fold-change of sorted to bulk
Top-hit #1	Intron 24/26 in BRCA2	6.75
Top-hit #2	Intron 2/26 in BRCA2	6.39
Top-hit #3	Upstream of N4BP2L1 gene	6.28

Note: The upper and lower tables rank the top-hit GFP^{neg} and GFP^{med} gRNAs with the highest log fold-change ratios, respectively.

Assessment of Controls and Sources of Bias in Enrichment-Based Estimations of gRNA Significance

The process of high-throughput gRNA screening to identify genomic sites of regulatory significance involves accounting for random variability in count data between populations, technical biases introduced by assay procedures and library preparation, and the uncertainty of measuring true gRNA abundance levels by read counts to confidently distinguish gRNAs that are associated with an expression loss phenotype (Anders & Huber, 2010; Rapaport et al., 2013). We calculate a log fold-change ratio between the sorted and bulk populations for each library gRNA, and prosecute a thresholding test based on the average log fold-change of the set of negative control gRNAs. Thus, a decision boundary line is formulated with the expectation that negative control gRNAs have a negative log fold-change ratio and that gRNAs of regulatory significance, as well as positive control gRNAs, will exceed the specified threshold by at least 2 standard deviations. In this section, we perform negative and positive control benchmarking to determine if the set of control gRNAs exhibit expected behaviors, and consequently diagnose the effects of low bulk read counts in skewing inferences on gRNA significance.

Analyzing Enrichment Scores of Negative and Positive Control gRNAs

The average log fold-change of the set of negative control, non-targeting gRNAs in GFP^{neg} cells is negative, indicating overall depletion of negative controls during purifying selection. At the level of individual negative control gRNAs, we observe that 55 gRNAs show no reads in the GFP^{neg} population but are present in the bulk population. 22 of the 71 negative control gRNAs represented in the GFP^{neg} population are enriched (log fold-change ratio > 0) - of the enriched negative controls, a strong majority (91%) are marginally enriched, but unexpectedly, 2 negative control gRNAs are significantly enriched in the GFP^{neg} population. It is necessary to scrutinize these two strongly enriched negative control gRNAs prior to further replicate screening in order to determine if their enrichment is caused by aberrant genome targeting that induces *BRCA2* expression loss.

The average log fold-change of the set of positive control gRNAs in GFP^{neg} cells is positive, upholding the expectation that these gRNAs have a stronger likelihood of inducing GFP loss. Individual analysis reveals that 52 of the 70 represented positive controls have a log fold-change ratio > 0, and unexpectedly, 4 positive control gRNAs are completely depleted in the GFP^{neg} population. While only 7 positive control gRNAs are significantly enriched above the 2 S.D. enrichment cutoff, we find that 29 gRNAs are enriched above a 1 S.D. enrichment cutoff in GFP^{neg} cells. Thus, the sets of negative and positive control gRNAs exhibit differential enrichment patterns, but 22/126 total negative control gRNAs (17%) and 22/74 total positive control gRNAs (30%) deviate from predicted behaviors. These results illustrate the essentiality of performing future biological and technical replicates to measure and filter out the background of random

biological variation and technical noise (Quackenbush, 2002). Additionally, the performance of control gRNAs is a useful touchstone to evaluate the reliability of the log fold-change metric and identify possible caveats that accompany using relative abundances to infer regulatory significance of genomic sites.

Effects of Low Read Counts on Enrichment Score Estimations

Log fold-change calculations are frequently employed in high-throughput sequencing assays to assess quantitative differences across sample states, such as differential expression comparisons of transcriptomics data or genome-wide analysis of regions preferentially associated with a molecular phenotype (Love, Huber, & Anders, 2014). We design an assay that enables quantification of log fold-change ratios across a large genomic terrain by exhaustive tiling of successive gRNAs such that drop-out of certain gRNAs (0 read count tally) in GFP^{neg} and GFP^{med} populations and system-wide variables such as randomized nucleotide insertions/deletions (indels) from end-joining repair can be tolerated in the analysis of a *BRCA2* functional regulatory architecture. However, while genome coverage is a necessary prerequisite for detecting a regulatory signal from read count data, it is also important to consider how low gRNA bulk counts can exaggerate log fold-change ratios to skew the signal. Figure 12 depicts a histogram of bulk read counts across the *BRCA2* gRNA library, highlighting a "low count" bin from 1 to 5 reads.

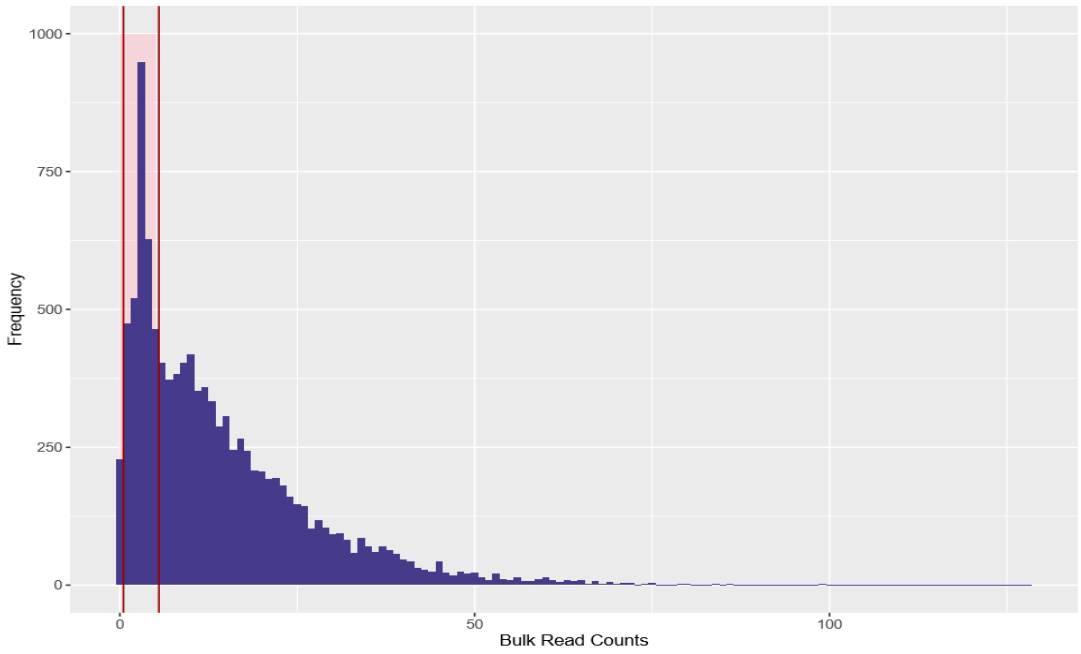

Figure 12. Histogram of *BRCA2* gRNA bulk read counts. The low count bulk read range (1-5 reads) is highlighted in pink.

The relationship between gRNA relative abundance and regulatory site importance is complicated by low bulk counts, which injects noisiness into enrichment ratio calculations. At low bulk count numbers (<= 5 reads), a single read difference between the bulk and sorted population has a magnified effect on the fold-change ratio, causing small changes to be construed as statistically significant above the enrichment score threshold. This "enrichment exaggeration" effect has a tendency to inflate the log fold-changes of gRNAs with low read counts while diluting the signal of gRNA candidates at the other end of the count spectrum – as such, it is possible that certain gRNAs, even if significantly enriched, may not be the most biologically relevant *cis*-regulatory sites for *BRCA2* transcription control (Love, Huber & Anders, 2014).

Modeling the log fold-change ratio of GFP^{neg} to bulk reads in relation to the bulk counts additionally illustrates a pattern of increased variance (heteroskedasticity) in log fold-changes depending on bulk count above the equatorial zero fold-change line, with numerous low count gRNAs landing above the enrichment score cutoff for significance (Figure 13).

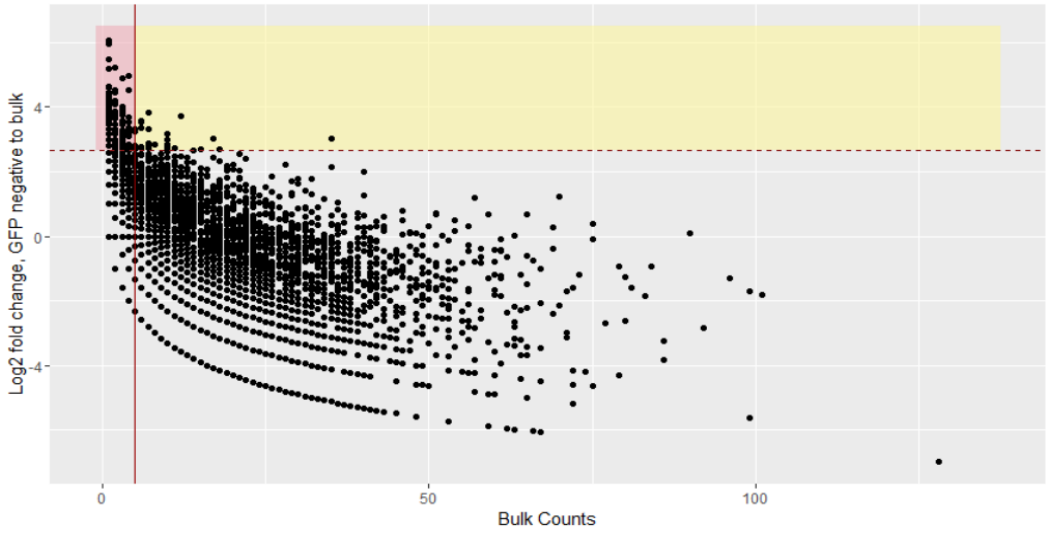

Figure 13. Log_2 fold-change of GFP^{neg} to bulk reads vs. bulk counts. Scatterplot of log-transformed fold change ratios between GFP^{neg} and bulk populations along bulk read counts. The horizontal dashed line indicates the enrichment score threshold for gRNA significance and the vertical full line indicates the boundary of the low bulk count bin (5 reads). Dividing the scatterplot into 4 quadrants, the upper left quadrant (pink) is the set of significantly enriched gRNAs with low bulk counts and the upper right quadrant (yellow) is the set of significantly enriched gRNAs with >5 bulk counts.

Determination of gRNA Significance by Absolute GFP^{neg} Counts

Here we evaluate gRNA significance to *BRCA2* expression by GFP^{neg} read count values, and then compare top-ranked gRNAs from an absolute count method of gRNA analysis to top-enriched gRNAs determined by the log fold-change metric. This method links gRNA abundance in GFP^{neg} cells to functional importance instead of using relative abundances, and thus does not rely on the sorted-to-bulk ratio to infer required elements for gene expression. First, we rank gRNAs based on absolute GFP^{neg} counts, observing a range of 0 to 285 reads. Out of 10576 library gRNAs, 4692 have reads >= 1 (GFP^{neg} represented gRNAs). Next, we assess negative and positive control performance: of the 126 negative control gRNAs, 55 have 0 GFP^{neg} reads, 24 have between 1 and 5 GFP^{neg} reads, and 13 have between 6 and 10 GFP^{neg} reads; of the 74 positive control gRNAs, 31 gRNAs are in the upper 10% of represented gRNAs. Conversely, 10 negative control gRNAs are present in the upper 10% of ranked GFP^{neg} gRNAs, and 4 positive control gRNAs have 0 GFP^{neg} counts and 12 have between 1 and 10 GFP^{neg} counts. By the absolute count method, 8% of negative control gRNAs and 22% of positive control gRNAs do not conform to expected depletion or enrichment behaviors, respectively. By control benchmarking between the two significance metrics, we find that the absolute GFP^{neg} count method exhibits improved negative and positive control performance.

Comparative analysis of top-ranked gRNAs by GFP^{neg} count and top-enriched gRNAs by log-fold change score demonstrates that the two methods can output divergent conclusions of significance for the same gRNA. Of the 10 maximal abundance gRNAs in the GFP^{neg} population, 6 do not reach significance by the enrichment score thresholding

method. The 2 *BRCA2*-targeting gRNAs that are considered significant by both criterions correspond to the promoter region of *N4BP2L1* and an intron within *FRY* (Table 2).

Table 2.

Top-ranked gRNAs by absolute GFP^{neg} read counts.

Top-ranked gRNAs by GFP^{neg} reads	Genome Location	Significance by Log-Fold Change Ratio
Top-rank #1	N4BP2L1 gene promoter	Yes
Top-rank #2	Intron 1/4 in N4BP2L1 gene	No
Top-rank #3	Intergenic region upstream of NBP2L1 promoter	No
Top-rank #4	Intron 51/60 in FRY gene	Yes
Top-rank #5	GM5 lncRNA	No
Top-rank #6	GFP-targeting positive control	Yes
Top-rank #7	GFP-targeting positive control	No
Top-rank #8	Intergenic region upstream of N4BP2L1 promoter	No
Top-rank #9	GFP-targeting positive control	Yes
Top-rank #10	Intergenic region downstream of N4BP212 gene	No

Note: Ranked gRNAs by maximum GFP^{neg} counts described by genome location and enrichment-based significance decision.

Interestingly, these 2 dual-significant gRNAs have high bulk read counts and fall below the upper bound of fold-change ratios, while the 3 most significantly enriched gRNAs possess low bulk reads and the highest log fold-change ratios. In future iterations of this assay, the two significance metrics should be applied in combination to enhance robust detection of functional *cis*-regulatory sites.

Detection of Significant Windows of Multiple Consecutive gRNAs

Here, we seek to determine the presence of significant regulatory regions in the GFP^{neg} population by integrating GFP^{neg} count information across multiple sequential gRNAs and performing a statistical test that compares the average GFP^{neg} count of a given gRNA window to the GFP^{neg} read count of the set of negative control gRNAs with a controlled FDR ≤ 0.05 (Fulco et al., 2016). This enables high-confidence identification of gRNA windows (~500 bp span) that exhibit strong association with expression loss compared to the distribution of non-targeting gRNA controls in the GFP^{neg} population, resulting in clarification of clustering dynamics of significant gRNAs. The output of the test is highly interpretable: if there are local clusters of over-represented gRNAs, then the corresponding window will be considered significant; if high abundance gRNAs are predominantly dispersed throughout the genome and do not concentrate in specific regions, then there will not be windows detected as significant. We do not conclude the presence of any significant regions, supporting previous analyses of a distributed regulatory architecture governing *BRCA2* expression. As such, even top-ranked gRNAs (maximal read counts) in the GFP^{neg} population do not reside in broader windows of regulatory significance.

Chapter IV

Discussion

The regulation of gene expression is highly interactive and dynamic, governed by TF motif recognition, cooperativity between TFs to stabilize binding and recruit additional factors to *cis*-modules, physical interaction between bound enhancers and the cognate promoter, and assembly of RNA Polymerase II and the full transcriptional machinery to initiate gene transcription (Wilczynski, Liu, Yeo, & Furlong, 2012). Various aspects of this regulatory choreography are not well elucidated, as it is challenging to quantify the regulatory activity of individual CREs in an endogenous context, discern the paired responsiveness between a gene-specific promoter and a TF-bound enhancer, and integrate activating cues across multiple CREs to construct a profile of necessary regulatory inputs for target gene expression (Cusanovich, Pavlovic, Pritchard, & Gilad, 2014; Wilczynski et al., 2012). Importantly, a lack of functional annotation in the regulatory genome impairs interpretation of regulatory sequence variations. Genome-wide association studies have uncovered thousands of non-coding variants associated with disease traits, but frequently are unable to distinguish silent alterations from deleterious mutations that affect functional regulatory sites, obfuscating the causal role of specific *cis*-regulatory variations in disease progression (Ward & Kellis, 2012).

Here we devise a CRISPR/Cas-based method to analyze the contribution of *cis*-regulatory DNA to gene expression. We find evidence that a spatially dispersed and functionally variegated set of genomic regulatory elements controls *BRCA2* expression, identifying several high-abundance gRNAs located >70 kb away from the *BRCA2* promoter region, deciphering regulatory participation of genomic domains with and without conventional chromatin features, and performing a "sliding window" statistical test to confirm the broad spread of high abundance gRNAs. We utilize two metrics to evaluate the significance of gRNA representation in GFP loss populations, enabling high-throughput recognition of *cis*-regulatory sites that are necessary for *BRCA2* expression.

This section discusses various approaches to validate the regulatory contribution of significant gRNAs, and future applications to dissect integrative aspects of gene control across multiple functional regions and analyze noncoding mutations that are causal to cancerous disease states.

Validation of Predicted Functional Regulatory Sites and Characterization of Off-target CRISPR/Cas Activity

Here we develop a proof-of-concept system to demonstrate the viability of a high-throughput, unbiased CRISPR/Cas methodology to systematically screen the regulatory genome surrounding *BRCA2*, and we expect to confirm experimental reproducibility by analyzing the correlations between biological and technical replicates. Upon execution of replicate screening, we recommend the following approaches to validate the regulatory significance of *cis*-sites governing *BRCA2* expression: first, apply refined statistical

methods to model variability of counts between replicates and correct inflated log-fold change estimates for low-count samples; second, assess the false-positive and false-negative rates of the pooled gRNA library format by testing individual gRNAs; and third, analyze the introduction of false positives from off-target effects of CRISPR/Cas cleavage.

Variability between replicates arises from biological heterogeneity in a cell population, technical variations from the sequencing process, and the randomized vocabulary of CRISPR-mediated indels at a given target site (Anders & Huber, 2010; Rajagopal et al., 2016). A major component of biological variation in the assay is the stochastic expression pattern of *BRCA2*, which is defined by random cellular transitions between active and inactive gene production. This "switch-like" property of expression can cause cells to shuttle in and out of the GFP^{neg} population independent of *cis*-element targeting, and some of these cells will be flow cytometrically preserved and sequenced in the GFP^{neg} pool. We observe that around 18% of the *BRCA2-GFP* cell population resides in an expression "off" state at any given time – although we do not have definitive means to discriminate between stochastic and gRNA-driven events of *BRCA2* attenuation, it is possible to filter out experimental noise from significance estimations by sharing information across numerous biological and technical replicates. Upon replicate collection, we suggest the application of statistical methodologies to model inter-replicate variability and execute controlled shrinkage of log fold-change ratios for low bulk counts prior to hypothesis testing of gRNA significance (Love, Huber & Anders, 2014). It is important to note that gRNA enrichments between biological replicates also vary as a consequence of unique mutant genotypes generated by CRISPR/Cas cleavage, and that

increasing the number of biological replicates will strengthen the detection of significant gRNAs (precision) above the levels of *BRCA2* expression variations, uncertainty of enrichment estimations, and randomized transmission of indels.

Assessment of false-positive and false-negative rates involves consideration of the experimental and computational liabilities associated with a high-throughput screening process that affect the accuracy of functional *cis*-regulatory identification. For the set of gRNAs categorized as significant across replicate testing, individual validation of induced expression loss in a non-pooled format will provide clarification of true positives and false positives. We expect that gRNAs detected as enriched by our assay will demonstrate comparable activity in smaller cell populations, and we can construct a receiver operating characteristic (ROC) curve to compare the true-positive rate against the false-positive rate. We can also determine the false-negative rate by computing the fraction of positive control gRNAs that induce GFP loss by individual follow-up tests and the proportion of validated positive control gRNAs that are identified as non-significant (false negatives) by our high-throughput system.

It is important to consider the potential for off-target CRISPR/Cas cleavage to affect precision of gRNA analysis and increase false discovery of functional regulatory sites. The specificity of Cas9 cleavage is governed by the targeting spacer sequence in the gRNA and the PAM recognition sequence in the genome, but the Cas9 nuclease can tolerate mismatches between the gRNA-DNA pairing, resulting in unwanted cleavage at non-target locations (Cho et al., 2014). To analyze false-positives caused by off-target effects, we recommend using off-target prediction tools (CRISPR Design, ZiFiT) to determine potential off-target sites of library gRNAs, and to eliminate gRNAs with off-

target scores from subsequent significance analysis (Marx, 2014). Ideally, a strong majority of significantly enriched or high-abundance gRNAs in the GFP^{neg} population will be retained and the spatial distribution of significant gRNAs will not be altered by off-target filtering.

Future Directions of Functional CRE Annotation in the Regulatory Genome

Functional maps of the regulatory genome can help elucidate the combinatorial regulation of gene expression patterns across multiple active *cis*-regulatory sites. Currently, little is known about how multiple regulatory regions integrate their activities to produce transcriptional responses that vary along spatial and temporal axes. Our system of CRE identification enables deeper analysis on the interrelationships and convergence of core regulatory elements that control gene expression through high-throughput introduction of specific gRNA pairs that can act synergistically, independently or in opposition to affect gene expression (Wilczynski et al., 2012). Having identified functional CREs that regulate a gene of interest such as *BRCA2*, we can also introduce targeted changes to the enhancer sequence that alter binding site multiplicity, TF site location and site order to address the roles of motif grammar and sequence context in determining the regulatory contribution of genomic elements (Sharon et al., 2012).

Our method of CRISPR/Cas-based perturbation of genomic regions can be harnessed to examine the mutant genotypes that are associated with an expression loss phenotype and describe novel sequence motifs that capacitate regulatory function. Cas9

cleavage of a targeted genomic site initiates an error-prone process of cellular repair at the specified element, resulting in a randomized process of genomic alteration that can range from a single nucleotide insertion to a 40-bp deletion (Marx, 2014). Our current pipeline assumes equivalence between the mutant genotypes that a single gRNA induces, as it does not distinguish between the distinct and variable regulatory mutations that a single gRNA produces, and only scores the functional significance of the gRNA based on its abundance in GFP^{neg} and GFP^{med} populations. However, our system can be extended to explore *cis*-regulatory sequence variations that cause expression loss with nucleotide-level resolution. Future experiments in this area involve individual introduction of select library gRNAs for targeted disruption, enumeration of the spectrum of mutant genotypes in GFP^{pos} and GFP^{neg} populations by deep sequencing, and algorithmic discovery of critical base positions (sequence motifs) within a given regulatory element and mutation types that abolish the element's activity (Rajagopal et al., 2016). The ultimate objective of these experiments is two-fold: first, to identify sequence features within a given required regulatory element that endow it with its functional activity; and second, to capture sequence variations along different positions of a regulatory motif that affect the functional capacity of the element (Ward & Kellis, 2012).

As an integral DNA repair gene with a high carrier rate and known association in breast and ovarian cancer susceptibility, *BRCA2* is a prime candidate for regulatory mapping. Approximately 1 in 240 individuals carry a pathogenic variant (germline or somatic mutation) of *BRCA1* or *BRCA2*, and *BRCA1*/*BRCA2* mutations account for 5-10% of breast cancers diagnosed in women under the age of 40 and explain ~24% of familial ovarian cancer cases (Jervis et al., 2014; Milne & Antoniou, 2016). *BRCA2*

genetic testing is offered through a number of companies (such as Myriad Genetics and Quest Diagnostics) in conjunction with a panel of other homologous recombination genes, but genetic screening is frequently limited to the coding frame regions and thus does not provide a comprehensive picture of breast and ovarian cancer risks (Walsh, 2015). Our system enables detailed annotation of functional enhancers and regulatory motifs with the discriminative power to distinguish between deleterious and passive mutations, and in the future, significant *BRCA2* elements discovered by our method can be decussated with genome-wide association studies of BRCA2-deficient cancers to assess the penetrance and pathogenicity of non-coding sequence modifications. Translation of functional regulatory maps to the clinic has the potential to extend the search coverage of cancer-associated mutations during screening, thus improving the accuracy of cancer risk diagnosis, guiding risk management strategies, and facilitating informed decision-making for treatment plans (Milne & Antoniou, 2016; Walsh, 2015).

Chapter V

Appendix

Table 3.

Primer sequences and experimental descriptions.

Primer Name	Primer Sequence	Primer Description
091514_U6gRNA_ROSAHDR_fw	CCAGGTTAGCCTTTAAGCCTGCCCAGAAGACTCCCGCCCA GCATGTGAGGGCCTATTTCC	PCR dummy gRNA plasmid with ROSA26 homology arms
091514_U6gRNA_ROSAHDR_rv	GGAGAATCCCTTCCCCCTCTTCCCTCGTGATCTGCA TCGCGATTTTACCACATTTGTAGA	PCR dummy gRNA plasmid with ROSA26 homology arms
091514_ROSAHDR_Ext_fw	ACACCTGTTCAATTCCCCTGCAGGACAACGCCCACACACCAG GTTAGCCTTTAAG CCTGC	Extend ROSA26 homology arms for dummy gRNA knock-in construction
091514_ROSAHDR_40bpext_rv	TCTGCTGCCTCCTGGCTTCTGAGGACCGCCCTGGGCCTGGGA GAATCCCTTCCC CCTCTT	Extend ROSA26 homology arms for dummy gRNA knock-in construction
080814_U6gRNA_late_fw	TCTACAAATGTGGTAAAATCGCGA	Genomic DNA PCR primers to verify dummy gRNA knock-in in clones
091514_ROSA_downstream_rv	GGGAGGGGAGTGTTGCAATA	Genomic DNA PCR primers to verify dummy gRNA knock-in in clones
091514_ROSA_upstream_fw	TGGGAAGTCTTGTCCCTCCA	Sequencing primers to sequence dummy gRNA cassette in clonal knock-in lines
091514_ROSA_downstream_rv	GGGAGGGGAGTGTTGCAATA	Sequencing primers to sequence dummy gRNA cassette in clonal knock-in lines
072715_sgBrca23_GFPHDR_fw	CTGTAGAGGCGACAGCAGTGAGAAATTAGCTGTTGAGTCT GTGAGCAAGGGCGAGGAGCT	PCR amplify GFP plasmid with BRCA2 homology arms
072715_sgBrca23_GFPHDR_rv	TTCTCACACGAACACCTATGAGTAGCCTGGAACTGTACAC TGAGGAGTGAATTGCGGCCG	PCR amplify GFP plasmid with BRCA2 homology arms
072715_sgBrca23_HDRext_fw	GCAAGTAGGGCCCAGGTCCAGGAAGGAGTCTCTCAGGGACT GTAGAGGCGACAGCAGTGA	Extend BRCA2 homology arms for GFP knock-in
072715_sgBrca23_HDRext_rv	ACACACGCTTCAGTAGAGTGCAGCTACTCCCGCCTTCTCACAC GAACACCTATGAGTAGC	Extend BRCA2 homology arms for GFP knock-in
072715_sgBrca23_up_fw	CGGTGATTCCACAAGGAAC	Genomic DNA PCR primers to verify GFP fusion with BRCA2
072715_sgBrca23_dwn_rv	CCAGGTAGAGCATCTGAGCA	Genomic DNA PCR primers to verify GFP fusion with BRCA2
052314_gRNALib_HDR_fw	TGTTTTAAAATGGACTATCATATGCTTACCGTAACTTGAAAGT ATTTCGATTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAA CACC	PCR amplify BRCA2 gRNA library with ROSA26 homology arms
101815_gRNALib_HDR_trunc_rv	CTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTT AAACTTGCTATGCTGTTTCCAGCATAGCTCTTAAAC	PCR amplify BRCA2 gRNA library with ROSA26 homology arms
082214_gRNA_upstream_fw	TTGTGGAAAGGACGAAACACC	Library prep PCR 1 primers
091514_ROSA_downstream_rv	GGGAGGGGAGTGTTGCAATA	Library prep PCR 1 primers
082214_gRNA_upstream_fw	TTGTGGAAAGGACGAAACACC	Library prep qPCR primers
020515_gRNA_qPCR_rv	GCCTTATTTAAACTTGCTATGCTGT	Library prep qPCR primers
101714_gRNAPE1_BcX	CTCTTTCCCTACACGACGCTCTTCCGATCTCCAATTTGTGGAA AGGACGAAACACC	Library prep PCR 2 primer to barcode bulk sample
101714_gRNAPE1_BcY	CTCTTTCCCTACACGACGCTCTTCCGATCTGCGTATTGTGGAA AGGACGAAACACC	Library prep PCR 2 primer to barcode GFP negative sample
101714_gRNAPE1_BcZ	CTCTTTCCCTACACGACGCTCTTCCGATCTTGAGC TTGTGGAAAGGACGAAACACC	Library prep PCR 2 primer to barcode GFP medium sample
010715_LibrarygRNA_PE2	CATTCCTGCTGAACCGCTCTTCCGATCTGCCTTATTTAAACTTG CTATGCTGT	Library prep PCR 2 primer (Illumina PE2)
100615_PE1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACG ACGCTCTTCCGATCT	Library prep PCR 3 primer (Illumina PE1)
100615_PE2	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGC TGAACCGCTCTTCCGATCT	Library prep PCR 3 primer (Illumina PE2)

References

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology,11.* doi: 10.1186/gb-2010-11-10-r106

Boeva, V. (2016). Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in eukaryotic cells. *Frontiers in Genetics, 7*. doi: 10.3389/fgene.2016.00024

Calo, E., & Wysocka, J. (2013). Modification of enhancer chromatin: what, how and why? *Molecular Cell*, *49*(5). doi: 10.1016/j.molcel.2013.01.038

Chen, J., Zhang, Z., Li, L., Chen, B.C., Revyakin, A., Hajj, B., …Liu, Z. (2014). Single-molecule dynamics of enhanceosome assembly in embryonic stem cells. *Cell, 156(6)*, 1274-1285. doi: 10.1016/j.cell.2014.01.062

Cho, S., Kim, S., Kim, Y., Kweon, J., Kim, H.S., Bae, S., & Kim, J. (2014). Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res, 24*(1), 132-141. doi: 10.1101/gr.162339.113

Erceg, J., et al. (2014). Subtle changes in motif positioning cause tissue-specific effects on robustness of an enhancer's activity. *PLOS Genetics, 10(1).* doi: 10.1371/journal.pgen.1004060

Farley, E.K., Olson, K.M., Zhang, W., Rokhsar, D.S., & Levine, M.S. (2016). Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. *PNAS, 113(23)*, 6508-6513. doi: 10.1073/pnas.1605085113

Fulco, C.P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S., Perez, E.M., …Engreitz, J.M. (2016). Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science, 354*(6313), 769-773. doi: 10.1126/science.aag2445

Gudmundsdottir, K., & Ashworth, A. (2006). The roles of BRCA1 and BRCA2 and associated proteins in the maintenance of genomic stability. *Oncogene*, *25*, 5864-5874. doi: 10.1038/sj.onc.1209874

He, X., Samee, A.H., & Blatti, C., & Sinha, S. (2010). Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative bindings and short-range repression. *PLOS Computational Biology, 6(9).* doi: 10.1371/journal.pcbi.1000935

Jervis, S., Song, H., Lee, A., Dicks, E., Tyrer, J., Harrington, P., …Antoniou, A.C. (2014). Ovarian cancer familial relative risks by tumor subtypes and by known ovarian cancer genetic susceptibility variants. *Journal of Medical Genetics, 51,* 108-113. doi: doi:10.1136/jmedgenet-2013-102015

Junion, G., Spivakov, M., Girardot, C., Braun, M., Gustafson, H., Birney, E., & Furlong, E. (2012). A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell, 148*, 473-486. doi: 10.1016/j.cell.2012.01.030

Liu, Z., et al. (2014). Enhancer activation requires trans-recruitment of a mega transcription factor complex. *Cell,* 159(2), 358-373. doi: 10.1016/j.cell.2014.08.027

Love, M.I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology, 15*. doi: 10.1186/s13059-014-0550-8

Maia, A., et al. (2012). Effects of BRCA2 cis-regulation in normal breast and cancer risk amongst BRCA2 mutation carriers. *Breast Cancer Res, 14(2).* doi: 10.1186/bcr3169

Marx, V. (2014). Gene editing: how to stay on target with CRISPR. *Nature Methods, 11*, 1021-1026. doi: 10.1038/nmeth.3108

Milne, R.L, & Antoniou, A.C. (2016). Modifiers of breast and ovarian cancer risks for BRCA1 and BRCA2 mutation carriers. *Endocr Relat Cancer, 10*. doi: 10.1530/ERC-16-0277

Neph, S., Stergachis, A.B., Reynolds, A., Sandstrom, R., Borenstein, E., & Stamatoyannopoulos, J.A. (2012). Circuitry and Dynamics of Human Transcription Factor Regulatory Networks. *Cell, 150,* 1274-1286. doi: 10.1016/j.cell.2012.04.040

Prakash, R., Zhang, Y., Feng, W., & Jasin, M. (2015). Homologous recombination and human health: the roles of BRCA1, BRCA2, and associated proteins. *Cold Spring Harb Perspect Biol, 7(4)*. doi: 10.1101/cshperspect.a016600

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nature Genetics, 32*, 496-501. doi: 10.1038/ng1032

Rajagopal, N., Srinivasan, S., Kooshesh, K., Guo, Y., Edwards, M.D., Banerjee, B., … Sherwood, R.I. (2016). High-throughput mapping of regulatory DNA. *Nature Biotechnology, 34,* 167-174. doi: 10.1038/nbt.3468

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., …Betel, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology, 14*. doi: 10.1186/gb-2013-14-9-r95

Sanjana, N.E., Wright, J., Zheng, K., Shalem, O., Fontanillas, P., Joung, J., …Zhang, F. (2016). High-resolution interrogation of functional elements in the noncoding genome. *Science, 353(6307),* 1545-1549. doi: 10.1126/science.aaf7613

Schmidt, H.G., Sewitz, S., Andrews, S.S., & Lipkow, K. (2014). An integrated model of transcription factor diffusion shows the importance of intersegmental transfer and quaternary protein structure for target site finding. *PLOS One, 9(10)*. doi: 10.1371/journal.pone.0108575

Sharon, E., Kalma, Y., Sharp, A., Sadka, T., Levo, M., Zeevi, D., …Segal, E. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotechnology, 30*(6), doi: 10.1038/nbt.2205

Spitz, F., & Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics, 13,* 613-626. doi: 10.1038/nrg3207

Todeschini, A., Georges, A., & Veitia, R.A. (2014). Transcription Factors: specific DNA binding and specific gene regulation. *Trends in Genetics, 30(6)*, 211-219. doi: 10.1016/j.tig.2014.04.002

Walsh, C.S. (2015). Two decades beyond BRCA1/2: Homologous recombination, hereditary cancer risk and a target for ovarian cancer therapy. *Gynecologic Oncology, 137,* 343-350. doi: 10.1016/j.ygyno.2015.02.017

Wang, T., Wei, J.J., Sabatini, D.M., & Lander, E.S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. *Science, 343(6166)*, 80-84. doi: 10.1126/science.1246981

Ward, L.D., & Kellis, M. (2012). Interpreting noncoding genetic variation in complex traits and human disease. *Nature Biotechnology, 30,* 1095-1106. doi: 10.1038/nbt.2422

Welcsh, P., & King, M. (2001). BRCA1 and BRCA2 and the genetics of breast and ovarian cancer. *Human Mol Genet, 10(7),* 705-713

Wilczynksi, B., Liu, Y., Yeo, Z., & Furlong, E.E.M. (2012). Predicting spatial and temporal gene expression using an integrative model of transcription factor occupancy and chromatin state. *PLoS Comput Biol, 8*(12). doi: 10.1371/journal.pcbi.1002798

Zhao, Y., Ruan, S., Pandey, M., & Stormo, G.D. (2012). Improved Models for Transcription factor binding site identification using nonindependent interactions. *Genetics, 191,* 781-790. doi: 10.1534/genetics.112.138685